

Additive Partially Linear Models for Massive Heterogeneous Data

Binhuan Wang

Department of Population Health
New York University School of Medicine

Yixin Fang*

Department of Mathematical Sciences
New Jersey Institute of Technology

Heng Lian

Department of Mathematics
City University of Hong Kong

Hua Liang

Department of Statistics
George Washington University

Abstract

We consider an additive partially linear framework for modelling massive heterogeneous data. The major goal is to extract multiple common features simultaneously across all sub-populations while exploring heterogeneity of each sub-population. This work generalizes the partially linear framework proposed in Zhao et al. (2016), which considers only one common feature. Motivated by Zhao et al. (2016), we propose an aggregation type of estimators for the commonality parameters that possess the asymptotic optimal bounds and the asymptotic distributions as if there were no heterogeneity. This oracle result holds when the number of sub-populations does not grow too fast and the tuning parameters are selected carefully. A plug-in estimator for the heterogeneity parameter is further constructed, and shown to possess the asymptotic distribution as if the commonality information were available. The performance of the proposed methods is evaluated via simulation studies and an application to the Medicare Provider Utilization and Payment data.

Key words: *Divide-and-conquer, heterogeneity, oracle property, regression splines*

1 Introduction

Recent revolutions in technologies have produced many kinds of massive data, where the number of variables p is fixed but the sample size N is very large. Wang et al. (2015) carried out a survey of statistical strategies for such data, and loosely grouped them into three categories: sub-sampling, divide and conquer, and sequential updating. Using the divide-and-conquer strategy, the original,

* Correspondence to: 210 Cullimore Hall, University Heights Newark, New Jersey 07102; Email: yixin.fang@njit.edu

full dataset is first split into manageable sub-datasets; the final result is then “averaged” from those individual results of the sub-datasets. Many methods based on the divide-and-conquer strategy have been developed for the analysis of massive homogeneous data. For example, Lin and Xi (2011) developed a computation and storage efficient algorithm for estimating equation estimation in massive data sets using the divide-and-conquer strategy. Chen and Xie (2014) applied the split-and-conquer strategy to generalized linear models and showed that it can substantially reduce computing time and computer memory requirements. In a more general framework, Li et al. (2013) studied the properties of the divide-and-conquer strategy when applied to any statistical inference problem in the analysis of massive homogeneous data.

However, the research is lacking for the analysis of massive heterogeneous data using the divide-and-conquer strategy, although the analysis of non-massive heterogeneous data has been well studied in the literature. For example, non-massive heterogeneous data can be handled by fitting mixture models (Aitkin and Rubin, 1985) and by modeling variance functions (Davidian and Carroll, 1987). As far as we are aware, Zhao et al. (2016) is the first paper, and the only paper, that considers the analysis of massive heterogeneous data using the divide-and-conquer strategy. In Zhao et al. (2016), they proposed a partially linear framework for modelling massive heterogeneous data, attempting to extract the common feature across all sub-populations while exploring heterogeneity of each sub-population. But the partially linear framework can only deal with only one common feature. In this paper, we propose an additive partially linear framework for modelling massive heterogeneous data, which can be applied to extract several common features across all sub-populations while exploring heterogeneity of each sub-population.

The additive partially linear models (APLMs) are a generalization of multiple linear regression models, and at the same time they are a special case of generalized additive nonparametric regression models (Hastie and Tibshirani, 1990). As discussed in Liu et al. (2011), APLMs allow an easier interpretation of the effect of each variable and are preferable to completely nonparametric additive models, since they combine both parametric and nonparametric components when it is believed that the response variable depends on some variables in a linear way but is nonlinearly

related to the remaining independent variables. Estimation and inference for APLMs have been well studied in literature (e.g., Carroll et al., 2003; Opsomer and Ruppert, 1999). Recently, Fang et al. (2015) proposed an approach for the analysis of heterogeneous data, fitting both the mean function and variance function using different additive partially linear models.

In this paper, we generalize the partially linear model (PLM) considered in Zhao et al. (2016) and propose an additive partially linear model (APLM) for modeling massive heterogeneous data. Let $\{(Y_i, \mathbf{X}_i, \mathbf{Z}_i)\}_{i=1}^N$ be the observations from a sample of N subjects. As in Zhao et al. (2016), we assume that there exist s independent sub-populations, and the data from the j th sub-population follow the following additive partially linear model,

$$Y^{(j)} = \mathbf{X}^T \boldsymbol{\beta}_0^{(j)} + \sum_{k=1}^K g_{0k}(Z_k) + \varepsilon, \quad (1)$$

where $\mathbf{X} = (X_1, \dots, X_d)^T$, $\mathbf{Z} = (Z_1, \dots, Z_K)$, $\boldsymbol{\beta}_0^{(j)} = (\beta_{01}^{(j)}, \dots, \beta_{0d}^{(j)})^T$ is the vector of unknown parameters for j th sub-population, g_{01}, \dots, g_{0K} are unknown smooth functions, and ε has zero mean and variance σ^2 . The partially linear model considered in Zhao et al. (2016) is a special case of (1) where $K = 1$.

Under model (1), $Y^{(j)}$ depends on \mathbf{X} linearly but with coefficients varying across different sub-populations, whereas $Y^{(j)}$ depends on \mathbf{Z} through additive nonlinear functions that are common to all sub-populations. This model implies that the heterogeneity of the data is coming from the difference among $\boldsymbol{\beta}_0^{(j)}$, $j = 1, \dots, s$. We revise the motivational scenario in Zhao et al. (2016) for our more general model (1): different labs conduct the same experiment on the relationship between a response variable $Y^{(j)}$ (say, heart disease) and a set of predictors \mathbf{X} and \mathbf{Z} . Prior knowledge shows that the relationship between $Y^{(j)}$ and \mathbf{Z} (say, systolic blood pressure (SBP), low-density lipoprotein cholesterol (LDL), and glycosylated hemoglobin (A1c)) should be homogeneous for all patients. However, the relationship between $Y^{(j)}$ and \mathbf{X} (say, certain genes) varies in different labs; for example, the genetic functionality of different races might be heterogenous.

The rest of the paper is organized as follows. We develop the methods and derive their asymp-

otic properties in Section 2. We evaluate the performance of the proposed methods via simulation studies in Section 3 and a real data application in Section 4. We conclude the paper with a brief summary in Section 5 and relegate all the technical proofs to the Appendix.

2 Methods

2.1 Notation and assumptions

Recall that $\beta_0^{(j)}$ is the true sub-population specific parameter-vector for the j th sub-population, $j = 1, \dots, s$, and $g_0(\mathbf{z}) = g_{01}(z_1) + \dots + g_{0K}(z_K)$ is the true additive common non-parametric function. Without loss of generality, assume that $g_{0k} = g_{0k}(\cdot)$, $k = 1, \dots, K$, have a common support $[0, 1]$. We propose to use polynomial splines (Carroll et al., 2003) to approximate smooth function g_{0k} , $k = 1, \dots, K$. Let \mathcal{S}_N be the space of polynomial splines on $[0, 1]$ of degree $\varrho \geq 1$, with a sequence of J_N interior knots,

$$t_{-\varrho} = \dots = t_{-1} = t_0 = 0 < t_1 < \dots < t_{J_N} < 1 = t_{J_N+1} = \dots t_{J_N+\varrho+1},$$

where J_N increases with the overall sample size N . Although we can choose different sequences of interior knots for different non-parametric functions in different sub-populations, for simplicity, as in Liu et al. (2011), here we consider the same sequence of equally spaced knots and let $h_N = 1/(J_N + 1)$ be the distance between neighboring knots.

Assume that \mathbf{X}_i are i.i.d. with \mathbf{X} and \mathbf{Z}_i are i.i.d. with \mathbf{Z} . Define $\mathbf{T} = (\mathbf{X}, \mathbf{Z})$. Let $m_0^{(j)}(\mathbf{T}) = \mathbf{X}^T \beta_0^{(j)} + g_0(\mathbf{Z})$, $\Gamma(\mathbf{z}) = E(\mathbf{X} | \mathbf{Z} = \mathbf{z})$, and $\widetilde{\mathbf{X}} = \mathbf{X} - \Gamma(\mathbf{Z})$. And $\mathbf{C}^{\otimes 2}$ denotes $\mathbf{C}\mathbf{C}^T$ for any matrix or vector \mathbf{C} . Let r be a positive integer and $\nu \in (0, 1]$ such that $p = r + \nu > 2$. Let \mathcal{H} be the collection of functions h on $[0, 1]$ whose r th derivative exists and satisfies the Lipschitz condition of order ν ,

$$|h^{(r)}(z') - h^{(r)}(z)| \leq C|z' - z|^\nu, \forall 0 \leq z', z \leq 1,$$

where and hereafter C is a generic positive constant. In order to derive asymptotic results, we make the following mild assumptions.

- (A1). Each component function $g_{0k} \in \mathcal{H}, k = 1, \dots, K$;
- (A2). The distribution of \mathbf{Z} is absolutely continuous and its density f is bounded away from zero and infinity on $[0, 1]^K$;
- (A3). The random vector \mathbf{X} satisfies that for any vector $\boldsymbol{\omega} \in \mathbb{R}^d$, $c\|\boldsymbol{\omega}\|^2 \leq \boldsymbol{\omega}^T E(\mathbf{X}^{\otimes 2} | \mathbf{Z} = \mathbf{z}) \boldsymbol{\omega} \leq C\|\boldsymbol{\omega}\|^2$, where c is a positive constant;
- (A4). The number of interior knots J_N satisfies: $N^{1/(4p)} \ll J_N \ll N^{1/4}$;
- (A5). The projection function $\Gamma(\mathbf{z})$ has the additive form $\Gamma(\mathbf{z}) = \Gamma_1(z_1) + \dots + \Gamma_K(z_K)$, where $\Gamma_k \in \mathcal{H}, E[\Gamma_k(z_k)] = 0$ and $E[\Gamma_k(z_k)]^2 < \infty, k = 1, \dots, K$.

In addition, to quantify the asymptotic consistencies of the non-parametric estimators, we consider both the empirical norms and the corresponding population norms. Let $\|\mathbf{z}\|$ be the Euclidean norm, $\|\mathbf{z}\|_\infty$ be the supremum norm, and $\|\mathbf{z}\|_1$ be the absolute-value norm of a vector \mathbf{z} , respectively. For a matrix \mathbf{C} , its L_2 -norm is defined as $\|\mathbf{C}\|_2 = \sup_{\|\mathbf{u}\| \neq 0} \|\mathbf{C}\mathbf{u}\|/\|\mathbf{u}\|$. Let $\|\varphi\|_\infty = \sup_{x \in [0, 1]} |\varphi(x)|$ be the supremum norm of a function φ on $[0, 1]$. Following Stone (1985) and Huang et al. (2003), for any measurable function ϕ_1 and ϕ_2 on $[0, 1]^K$, the empirical inner product and norm for the j th sub-sample and the whole sample, respectively, are defined as

$$\begin{aligned} \langle \phi_1, \phi_2 \rangle_{jn} &= \frac{1}{n} \sum_{i \in \mathcal{G}_j} \phi_1(\mathbf{Z}_i) \phi_2(\mathbf{Z}_i), \quad \|\phi\|_{jn}^2 = \frac{1}{n} \sum_{i \in \mathcal{G}_j} \phi^2(\mathbf{Z}_i), \\ \langle \phi_1, \phi_2 \rangle_N &= \frac{1}{N} \sum_{i=1}^N \phi_1(\mathbf{Z}_i) \phi_2(\mathbf{Z}_i), \quad \|\phi\|_N^2 = \frac{1}{N} \sum_{i=1}^N \phi^2(\mathbf{Z}_i), \end{aligned}$$

If ϕ_1 and ϕ_2 are L^2 -integrable, the population inner product and norm are defined as

$$\langle \phi_1, \phi_2 \rangle = \int_{[0, 1]^K} \phi_1(\mathbf{z}) \phi_2(\mathbf{z}) f(\mathbf{z}) d\mathbf{z}, \quad \|\phi\|_2^2 = \int_{[0, 1]^K} \phi^2(\mathbf{z}) f(\mathbf{z}) d\mathbf{z},$$

where f is the density of \mathbf{Z} . Similarly, for the k th component of \mathbf{Z} , Z_k with density f_k , the empirical norm on the j th sub-sample, the empirical norm on the whole sample, and the population norm of any L^2 -integrable univariate function φ on $[0, 1]$ are defined as

$$\|\varphi\|_{jnk}^2 = \frac{1}{n} \sum_{i \in \mathcal{G}_j} \varphi^2(Z_{ik}), \quad \|\varphi\|_{Nk}^2 = \frac{1}{n} \sum_{i=1}^N \varphi^2(Z_{ik}), \quad \|\varphi\|_{2k}^2 = \int_0^1 \varphi^2(z_k) f_k(z_k) dz_k.$$

2.2 Estimations for each sub-population

First we consider the estimations for $\beta_0^{(j)}$ and $g_0 = g_0(\cdot)$ based on the data from the j th sub-population only, $j = 1, \dots, s$. To this aim, let G_j denotes the index set of all the observations from the sub-population j , and let $\mathcal{G}_n^{(j)} = \{g^{(j)}(\cdot)\}$ be the collection of additive functions with the form that $g^{(j)}(\mathbf{z}) = g_1^{(j)}(z_1) + \dots + g_K^{(j)}(z_K)$, where each component function $g_k^{(j)} \in \mathcal{S}_N$ and $\sum_{i \in G_j} g_k^{(j)}(Z_{ik}) = 0$. Thus $\sum_{i \in G_j} g^{(j)}(\mathbf{Z}_i) = 0$ for any $g^{(j)} \in \mathcal{G}_n^{(j)}$. For the j th sub-population, we consider the following estimators,

$$(\hat{\beta}^{(j)}, \hat{g}^{(j)}) = \underset{\beta \in \mathbb{R}^d, g \in \mathcal{G}_n^{(j)}}{\operatorname{argmin}} \left\{ L_n^{(j)}(\beta, g) = \frac{1}{2} \sum_{i \in G_j} [Y_i - \mathbf{X}_i^T \beta - g(\mathbf{Z}_i)]^2 \right\}. \quad (2)$$

For the k th covariate Z_k , let $b_{m,k}(z_k)$ be the B-spline basis functions of degree ϱ equipped with J_N knots defined above. For any $g \in \mathcal{G}_n^{(j)}$, we can write $g(\mathbf{z}) = \mathbf{b}(\mathbf{z})^T \boldsymbol{\gamma}$, where $\mathbf{b}(\mathbf{z}) = \{b_{m,k}(z_k), m = -\varrho, \dots, J_N, k = 1, \dots, K\}^T$, which is a $K(J_N + \varrho + 1)$ -dim vector given \mathbf{z} , along with $K(J_N + \varrho + 1)$ -dim coefficient-vector $\boldsymbol{\gamma} = \{\gamma_{m,k}, m = -\varrho, \dots, J_N, k = 1, \dots, K\}^T$. Therefore, (2.4) is equivalent to

$$(\hat{\beta}^{(j)}, \hat{\boldsymbol{\gamma}}^{(j)}) = \underset{\beta \in \mathbb{R}^d, \boldsymbol{\gamma} \in \mathbb{R}^{K(J_N + \varrho + 1)}}{\operatorname{argmin}} \left\{ l_n^{(j)}(\beta, \boldsymbol{\gamma}) = \frac{1}{2} \sum_{i \in G_j} [Y_i - \mathbf{X}_i^T \beta - \mathbf{b}(\mathbf{Z}_i)^T \boldsymbol{\gamma}]^2 \right\}, \quad (3)$$

if we consider the empirically centered estimator $\hat{g}^{(j)}(\mathbf{z}) = \sum_{k=1}^K \hat{g}_k^{(j)}(\mathbf{z})$, where

$$\hat{g}_k^{(j)}(z_k) = \sum_{m=-\varrho}^{J_N} \hat{\gamma}_{m,k} b_{m,k}(z_k) - \frac{1}{n} \sum_{i \in G_j} \sum_{m=-\varrho}^{J_N} \hat{\gamma}_{m,k} b_{m,k}(z_{ik}). \quad (4)$$

We derive some asymptotic results associated with the sub-population specific estimators, summarized in the following theorem. All the technical proofs are relegated to the Appendix.

Theorem 1 *Under Assumptions (A1)-(A5), if the number of knots satisfies that $J_N \ll n^{1/2}$, we have, for each sub-population, $j = 1, \dots, s$,*

$$\begin{aligned} \|\hat{g}^{(j)} - g_0\|_2 &= O_P \left(J_N^{1/2} n^{-1/2} + h_N^p \right) \\ \text{and } \|\hat{g}^{(j)} - g_0\|_{jn} &= O_P \left(J_N^{1/2} n^{-1/2} + h_N^p \right). \end{aligned}$$

If the number of knots further satisfies that $J_N \gg n^{1/(2p)}$ we have

$$\sqrt{n}(\hat{\beta}^{(j)} - \beta_0^{(j)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{D}^{-1}),$$

where $\mathbf{D} = E(\tilde{\mathbf{X}}^{\otimes 2})$.

Remark 1: Assume that we consider $s = O(N^{1-\gamma})$ sub-samples, each sub-sample of $n = O(N^\gamma)$ observations, where γ is some positive number between 0 and 1. In order to minimize the mean-square error of estimating g_0 , $O_P(J_N^{1/2} n^{-1/2} + h_N^p)$, the best selection of J_N is $O(N^{\frac{\gamma}{2p+1}})$, or equivalently, $O(n^{\frac{1}{2p+1}})$. Under this selection, the mean-square error achieves the optimal rate, $O(N^{\frac{p\gamma}{2p+1}})$, or equivalently, $O(n^{\frac{p}{2p+1}})$.

Remark 2: On the other hand, in order to ensure that $\hat{\beta}^{(j)}$ is \sqrt{n} -consistent for estimating $\beta_0^{(j)}$, we should adopt under-smoothing tuning with $J_N \gg n^{1/(2p)}$ and carefully determine a balance between the number of sub-samples and the size of each sub-sample. For example, this can be achieved if we select J_N as $O(N^q)$ with $1/(4p) < q < 1/4$, and consider $s = O(N^{1-\gamma})$ sub-samples, each

sub-sample of $n = O(N^\gamma)$, with $2q < \gamma < 2pq$. The order of J_N is consistent with the existing results in the literature. The recommended balance between s and n provides a guidance for the appropriate application of the divide-and-conquer strategy.

2.3 Aggregation of commonality

We consider the aggregated estimator, $\bar{g}(z) = \frac{1}{s} \sum_{j=1}^s \hat{g}^{(j)}(z)$, as the final estimator of $g_0(z)$ based on the whole sample. Let \mathcal{G}_N be the collection of functions with the additive form $g(z) = g_1(z_1) + \dots + g_K(z_K)$, where $g_k \in \mathcal{S}_N$ and $\sum_{j=1}^s \sum_{i \in G_j} g_k(Z_{ik}) = 0$. Thus, for any $g \in \mathcal{G}_N$, $\sum_{j=1}^s \sum_{i \in G_j} g(Z_i) = 0$. In order to ensure that $\bar{g} \in \mathcal{G}_N$, as in (4), we center the individual estimator $\hat{g}_k^{(j)}(z_k)$ via $\hat{g}_k^{(j)}(z_k) = \sum_{m=-\varrho}^{J_N} \hat{\gamma}_{m,k} b_{m,k}(z_k) - \frac{1}{N} \sum_{i=1}^N \sum_{m=-\varrho}^{J_N} \hat{\gamma}_{m,k} b_{m,k}(z_{ik})$. To abuse the notation, we still denote the centered estimator as $\hat{g}_k^{(j)}(z_k)$ and $\hat{g}^{(j)}(z) = \sum_{k=1}^K \hat{g}_k^{(j)}(z_k)$. We derive the mean-square error of \bar{g} in the following theorem.

Theorem 2 *Under Assumptions (A1)-(A5), if $J_N \ll n^{1/2}$, we have*

$$\|\bar{g} - g_0\|_2 = O_P \left(J_N^{1/2} N^{-1/2} + h_N^p \right), \text{ and } \|\bar{g} - g_0\|_N = O_P \left(J_N^{1/2} N^{-1/2} + h_N^p \right).$$

Remark 3: In order to minimize the mean-square error of estimating g_0 using the aggregated estimator, if we select J_N as $O(N^{\frac{1}{2p+1}})$, the mean-square error achieves the optimal rate $O(N^{\frac{p}{2p+1}})$.

Remark 4: We compare the mean-square error of \bar{g} with that of the following ‘‘oracle estimator’’:

$$\hat{g}_{\text{oracle}} = \underset{g \in \mathcal{G}_N}{\operatorname{argmin}} \frac{1}{2} \sum_{j=1}^s \sum_{i \in G_j} \left[Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_0^{(j)} - g(\mathbf{Z}_i) \right]^2.$$

assuming $\boldsymbol{\beta}_0^{(j)}$, $j = 1, \dots, s$, are known. Following the proof of Theorem 1, we can show that $\|\hat{g}_{\text{oracle}} - g_0\|_2 = O_P \left(J_N^{1/2} N^{-1/2} + h_N^p \right)$. Therefore, as long as $n \gg J_N^2$, the means-square errors of the aggregated estimator \bar{g} and the oracle estimator \hat{g}_{oracle} are of the same order.

We conclude this subsection with some results for the massive homogeneous data where $\boldsymbol{\beta}_0^{(j)} \equiv \boldsymbol{\beta}_0$, $j = 1, \dots, s$. These results are of their own interest, when the divide-and-conquer strategy

is applied to massive homogeneous data, where β_0 and g_0 are estimated using the aggregated estimators $\bar{\beta} = \frac{1}{s} \sum_{j=1}^s \hat{\beta}^{(j)}$ and \bar{g} , respectively. The result for \bar{g} is the same as that in Theorem 2 and the result for $\bar{\beta}$ is stated in the following corollary.

Corollary 1 *Consider homogeneous massive data where $\beta_0^{(j)} \equiv \beta_0, j = 1, \dots, s$. Under Assumptions (A1)-(A5), if $J_N \gg N^{1/(2p)}$ and $n \gg N^{1/2}$, we have*

$$\sqrt{N}(\bar{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{D}^{-1}).$$

2.4 Efficiency boosting for heterogeneous parameters

The asymptotic variance matrix of $\hat{\beta}^{(j)}$ derived in Theorem 1 shows that there is some room to improve the estimation efficiency, because $\mathbf{D}^{-1} = E^{-1}(\tilde{\mathbf{X}}^{\otimes 2})$ is bigger than the Cramer-Rao lower bound, $E^{-1}(\mathbf{X}^{\otimes 2})$. Therefore, we re-substitute the aggregated estimator of g , \bar{g} , into to improve the efficiency of estimating $\beta_0^{(j)}$. This leads to the following more efficient estimator,

$$\check{\beta}^{(j)} = \underset{\beta^{(j)} \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2} \sum_{i \in G_j} \left[Y_i - \mathbf{X}_i^T \beta^{(j)} - \bar{g}(\mathbf{Z}_i) \right]^2. \quad (5)$$

for $j = 1, \dots, s$. We derive the asymptotic normality of $\check{\beta}^{(j)}$ in the following theorem.

Theorem 3 *Under Assumptions (A1)-(A5), if J_N satisfies the condition that $J_N \ll n^{1/2}$ given in the first part of Theorem 1 and the condition that $J_N \gg N^{1/(2p)}$ given in Corollary 1, and it further satisfies that $J_N \ll s^{1/2}$, then we have*

$$\sqrt{n}(\check{\beta}^{(j)} - \beta_0^{(j)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{A}^{-1}),$$

where $\mathbf{A} = E(\mathbf{X}^{\otimes 2})$.

Remark 5: As in Remarks 1-2, assume that we consider $s = O(N^{1-\gamma})$ sub-samples, each sub-sample of $n = O(N^\gamma)$ observations, where γ is some positive number between 0 and 1. In order

to satisfy all the conditions in Theorem 3, we can consider $N^{2q} \ll n \ll N^{1-2q}$, with $1/(2p) < q < 1/4$, and select $J_N = O(N^q)$. If \mathbf{X} and \mathbf{Z} are not independent, $\mathbf{A}^{-1} < \mathbf{D}^{-1}$. But, in order to achieve such efficiency boosting, there are more conditions on the balance between n and s .

2.5 Testing heterogeneity

As in Zhao et al. (2016), we also develop statistical tests for the heterogeneity across sub-populations. For this aim, consider the following general class of pairwise testing hypotheses for heterogeneous parameters:

$$H_0 : \mathbf{Q}(\beta_0^{(j_1)} - \beta_0^{(j_2)}) = \mathbf{0}, \quad (6)$$

where $j_1 \neq j_2 \in \{1, \dots, s\}$, and $\mathbf{Q} = (\mathbf{q}_1^T, \dots, \mathbf{q}_{d_1}^T)^T$ is a $d_1 \times d$ matrix with $d_1 \leq d$. This class of tests includes testing if either the whole vector or specific entries of $\beta_0^{(j_1)}$ are equal to those of $\beta_0^{(j_2)}$. It is straightforward to construct two test statistics as follows,

$$\mathbf{Q}(\hat{\beta}^{(j_1)} - \hat{\beta}^{(j_2)}), \text{ or } \mathbf{Q}(\check{\beta}^{(j_1)} - \check{\beta}^{(j_2)}),$$

which are based on the estimators from Subsection 2.2 or the estimators from Subsection 2.4, respectively. We summarize the asymptotic properties of these two test statistics in the following theorem, based on which we can conduct Wald tests.

Theorem 4 *If the conditions in Theorem 1 hold, under the null hypothesis (6), we have*

$$\sqrt{n}\mathbf{Q}(\hat{\beta}^{(j_1)} - \hat{\beta}^{(j_2)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, 2\sigma^2 \mathbf{Q}\mathbf{D}^{-1}\mathbf{Q}^T).$$

Furthermore, if the conditions in Theorem 3 hold, under the null hypothesis (6), we have

$$\sqrt{n}\mathbf{Q}(\check{\beta}^{(j_1)} - \check{\beta}^{(j_2)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, 2\sigma^2 \mathbf{Q}\mathbf{A}^{-1}\mathbf{Q}^T).$$

3 Simulation Studies

We conduct simulation studies to examine the impact of the balance between sub-population sizes n and the number of sub-population s on the performance of the proposed estimators, \bar{g} and $\check{\beta}^{(j)}$. We consider the following additive partially linear model with two nonparametric components ($K = 2$) as the data generating model:

$$\begin{aligned} Y^{(j)} &= X\beta_0^{(j)} + g_1(Z_1) + g_2(Z_2) + \varepsilon, \\ g_1(Z_1) &= 5 \sin\{2\pi(Z_1 + 1)\}, \\ g_2(Z_2) &= 100 \left(e^{-1.625(Z_2+1)} - 4e^{-3.25(Z_2+1)} + 3e^{-4.825(Z_2+1)} \right) - C_0, \end{aligned}$$

where ε is generated from normal distribution $N(0, 1)$, Z_1, Z_2 and W are generated independently from uniform distribution $U(-1, 1)$, $X = \frac{1}{2}(W + Z_1)$, and C_0 is taken as $100(1 - e^{-3.25})/3.25 - 400(1 - e^{-6.5})/6.5 + 300(1 - e^{-9.75})/9.75$ to make sure that $E\{g_1(Z_1)\} = E\{g_2(Z_2)\} = 0$. We can show that $\tilde{X} = W/2$, $\mathbf{D} = E(\tilde{X}^2) = 1/12$, and $\mathbf{A} = E(X^2) = 1/6$. In order to generate heterogenous data, we let $\beta_0^{(j)} = j$, for the j th sub-population, $j = 1, \dots, s$, with $d = 1$.

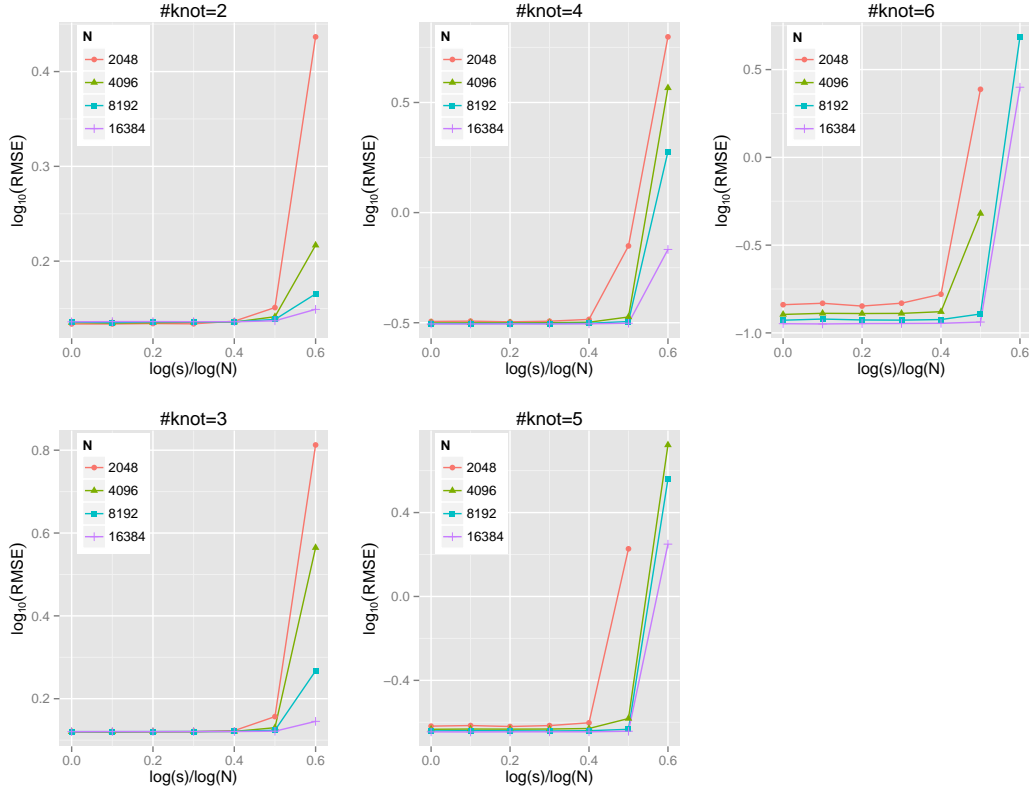
In order to g_1 and g_2 using polynomial splines, we consider cubic splines ($\varrho = 3$) and equal-spaced knots. We estimate the unknown error variance σ^2 using $\bar{\sigma}^2 = \sum_{j=1}^s (\hat{\sigma}^{(j)})^2 / s$, where

$$(\hat{\sigma}^{(j)})^2 = \frac{1}{n - d - K(J_N + \varrho)} \sum_{i \in G_j} \left[Y_i - X_i \hat{\beta}^{(j)} - \hat{g}^{(j)}(\mathbf{Z}_i) \right]^2.$$

We set the massive sample size N as $2^{11}, 2^{12}, 2^{13}$, or 2^{14} . We set the number of sub-samples s as $N^{1-\gamma}$, where $\gamma = \max(0.4, 2q), \dots, 0.9, 1$. We set the minimal value of γ as $\max(0.4, 2q)$ to ensure that $J_N^2 = O(N^{2q}) \ll n = O(N^\gamma)$. For each setting, we run 200 repetitions.

First, we evaluate the performance of the aggregated estimator, \bar{g} , as an estimator for g . We compute the root mean-square-error (RMSE) of \bar{g} , under different choices of J_N and s , and different settings of N . The results are summarized in Figure 1. The condition that $J_N^2 \ll n$, which is needed in all the theorems, implies that the larger number of knots we take and the shorter range

Figure 1: Root mean-square-errors of the aggregated estimator, \bar{g} , under different settings of the number of knots, the number of sub-samples, and the sample size.

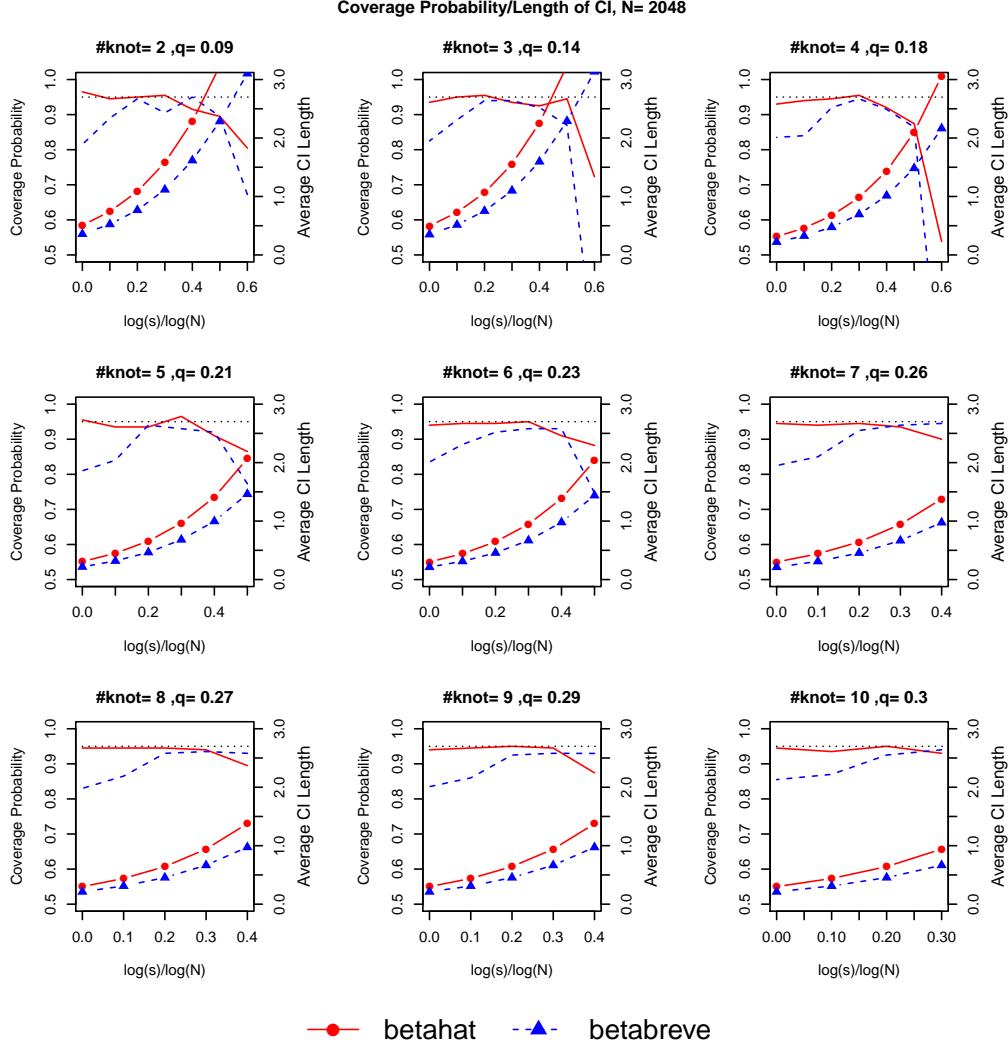


of s we should consider. In Figure 1, for each selection of the number of knots, we see that the performance of \bar{g} is good and stable during a wide range of s . We also see that the RMSE of \bar{g} deteriorates quickly when $\log(s)/\log(N)$ is approaching $1 - 2q$, $q \approx \log_N(J_N)$. For example, using 5 knots, $N = 2^{11}$, $q = \log_N(5) \approx 0.21$ and then $1 - 2q \approx 0.42$; therefore, from the second figure in the bottom row of Figure 1, we see that corresponding RMSE increases a lot when the ratio approaches 0.5. In summary, from 1, we see there is a clear boundary of $\log(s)/\log(N)$: with this boundary, the performance of \bar{g} is very good, while beyond this boundary, the performance is very bad. These findings confirm the theoretical results presented in Theorem 2.

Second, we evaluate the performance of the proposed estimators, $\hat{\beta}^{(j)}$ and $\check{\beta}^{(j)}$, for estimating $\beta_0^{(j)}$. We consider 95% confidence intervals based on $\hat{\beta}^{(j)}$ and $\check{\beta}^{(j)}$ respectively as follows:

$$\text{CI}_1 = \left[\hat{\beta}^{(j)} \pm \frac{1.96\bar{\sigma}}{\sqrt{n}} \mathbf{D}^{-1/2} \right] \quad \text{and} \quad \text{CI}_2 = \left[\check{\beta}^{(j)} \pm \frac{1.96\bar{\sigma}}{\sqrt{n}} \mathbf{A}^{-1/2} \right].$$

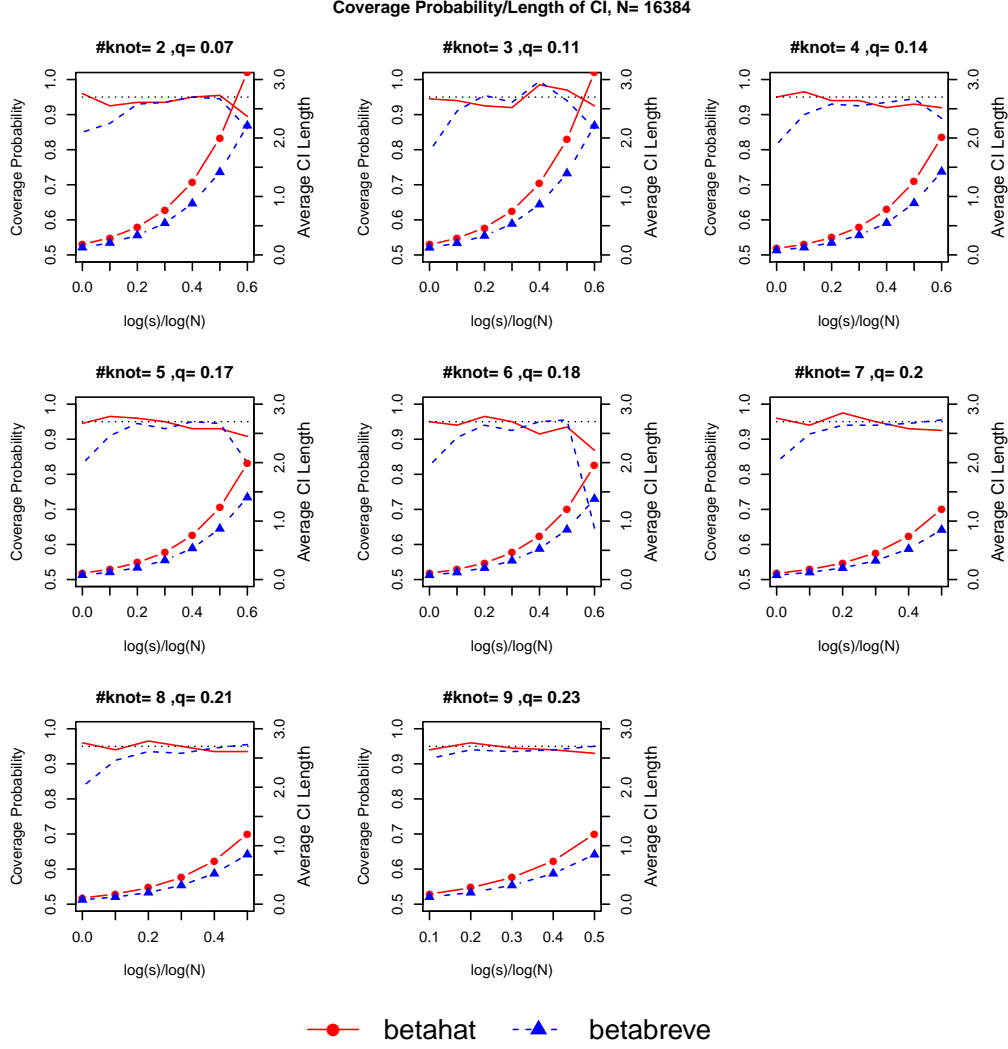
Figure 2: Coverage probabilities and interval lengths of 95% confidence intervals, CI_1 and CI_2 , under different settings of the number of knots and the number of sub-samples, with $N = 2^{11}$.



For simplicity, we summarize results for the first sub-population in Figures 2-4, where both the coverage probabilities and the interval lengths are displayed, with the results of $\hat{\beta}^{(1)}$ in red line with circle and those of $\breve{\beta}^{(1)}$ in blue dashed line with triangle.

From Figure 2 where $N = 2^{11}$ and Figure 3 where $N = 2^{14}$, we see that within a proper range of s , CI_1 and CI_2 have similar coverage probabilities. We also see that on average, the interval length of CI_2 is shorter than that CI_1 . This finding confirm that the asymptotic variance derived in Theorem 3 is smaller than that in Theorem 1. However, the coverage probability of CI_2 is valid for a shorter range of $\log(s)/\log(N)$, in contrast with that of CI_1 . This is finding is consistent with

Figure 3: Coverage probabilities and interval lengths of 95% confidence intervals, CI_1 and CI_2 , under different settings of the number of knots and the number of sub-samples, with $N = 2^{14}$.

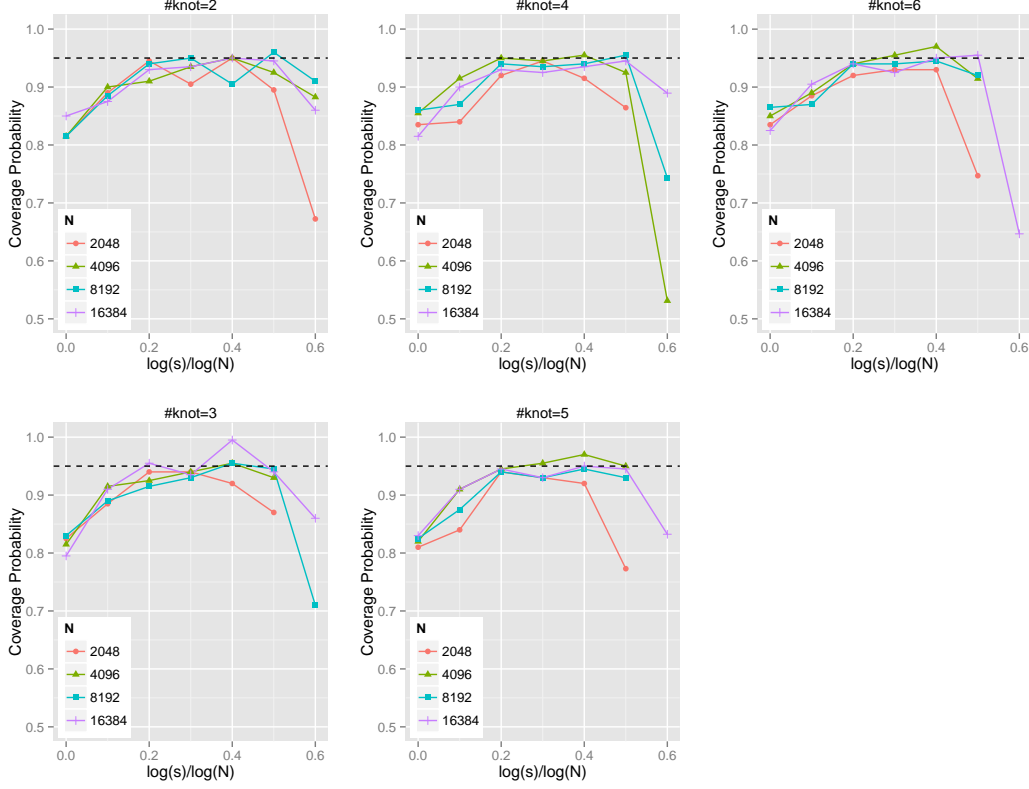


that there are more conditions in Theorem 3 than in Theorem 1.

To visualize the performance of CI_2 more clearly, in Figure 4 we display the coverage probability of CI_2 in more detail for different settings of s and N , given different numbers of knots. From Figure 4, we can see that, given the number of knots, a larger N implies a wider valid range for s to achieve a good coverage; given N , a larger number of knots implies a smaller transition point for s .

Third, we evaluate the heterogeneity tests using the following Wald test statistics constructed

Figure 4: Coverage probabilities of 95% CI_2 confidence intervals under different settings of the number of knots, the number of sub-samples, and the sample size.



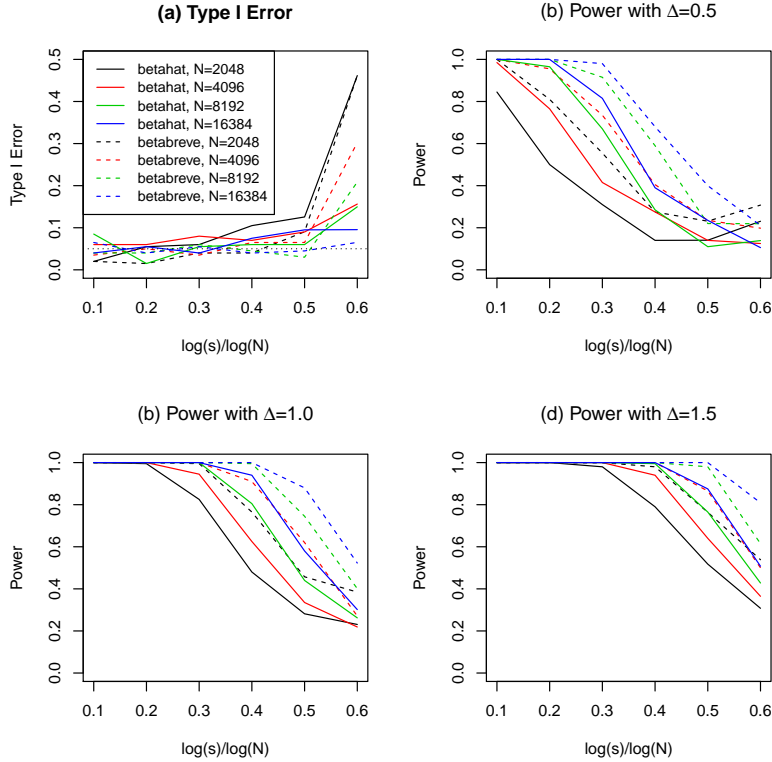
based on Theorem 4:

$$\begin{aligned}\Psi_1 &= I \left\{ \mathbf{Q}(\hat{\beta}^{(j_1)} - \hat{\beta}^{(j_2)}) \notin \sqrt{\frac{2}{n}} \bar{\sigma} (\mathbf{Q} \hat{\mathbf{D}}^{-1} \mathbf{Q}^T)^{1/2} C_{\alpha/2} \right\}, \\ \Psi_2 &= I \left\{ \mathbf{Q}(\check{\beta}^{(j_1)} - \check{\beta}^{(j_2)}) \notin \sqrt{\frac{2}{n}} \bar{\sigma} (\mathbf{Q} \hat{\mathbf{A}}^{-1} \mathbf{Q}^T)^{1/2} C_{\alpha/2} \right\},\end{aligned}$$

where $C_{\alpha/2}$ is the upper $\alpha/2$ quantile of a standard normal distribution, and $\hat{\mathbf{D}}$ and $\hat{\mathbf{A}}$ are the sample estimators of \mathbf{D} and \mathbf{A} , respectively. The results are summarized in Figure 5, where Ψ_1 and Ψ_2 are compared in terms of Type-I error and power, under different settings of s and N . From Panel (a) of Figure 5, we see that both Ψ_1 and Ψ_2 have appropriate type-I error within a wide range of s , but they have inflated type-I error after s passes a transition point. Panels (b)-(d) compare the testing powers under three different alternative hypotheses: $H_1 : \beta_0^{(j_1)} - \beta_0^{(j_2)} = \Delta$, where $\Delta = 0.5, 1$

and 1.5, respectively. We see that the power increases as N increase and Δ increases. We also see the power of Ψ_2 is larger than that of Ψ_1 across different settings. These findings confirm the asymptotic results stated in Theorem 4.

Figure 5: Type-I error and power of tests Ψ_1 and Ψ_2 under different settings of the number of sub-samples and the sample size, using 4 knots.



4 Real data application

We apply the proposed divide-and-conquer strategy for APLMs to the Medicare Provider Utilization and Payment Data (the Physician and Other Supplier Public Use File), with information on services and procedures provided to Medicare beneficiaries by physicians and other healthcare professionals. This dataset was prepared by the Centers for Medicare & Medicaid Services (CMS), as part of the Obama Administrations efforts to make our healthcare system more transparent, affordable, and accountable. We downloaded the dataset “Medicare Physician and Other Supplier Data

CY 2014” from `www.CMS.gov` with more than nine million records for health care providers from the U.S. or U.S. possessions. We focus on the subset consisting of 50 U.S. states and the District of Columbia (DC), which account for the majority part of the whole dataset.

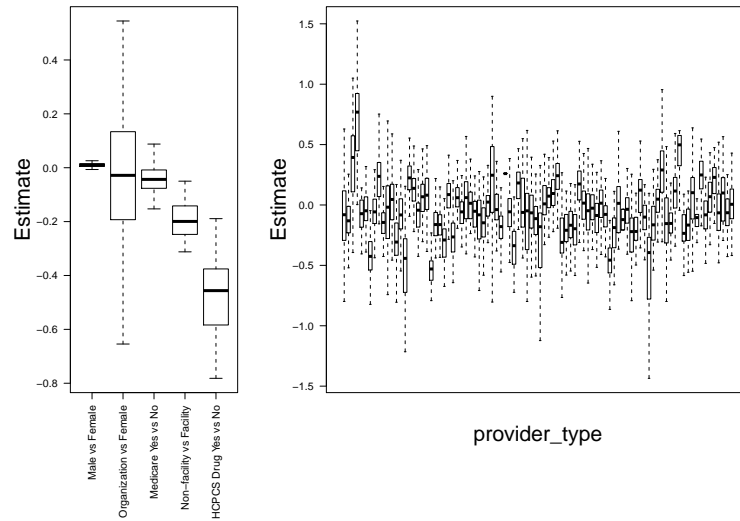
Our goal is to model the outcome variable “average_Medicare_standardized_amt” (average amount that Medicare paid after beneficiary deductible and coinsurance amounts have been deducted for the line item service and after standardization of the Medicare payment has been applied) on other covariates, including gender or entity of provider, provider type, Medicare participation status, place of service, HCPCS drug indicator, number of distinct Medicare beneficiaries (“bene_unique_cnt”), number of services provided (“bene_day_srvs_cnt”), and number of distinct Medicare beneficiary/per day services (“line_srvs_cnt”). Detailed explanations of these variables can be found in the official website `www.CMS.gov`. All covariates except the last three are categorical variables, and particularly the variable for provider type has 91 categories. Because those three quantitative variables are all count data, we take the \log_{10} -transformation and rescale each of them to the range $[-1, 1]$ by using the formula $(Z - \min Z)/(\max Z - \min Z) \times 2 - 1$. Also, we apply the \log_{10} -transformation to the outcome variable, which is skewed to the right. By excluding those records with value 0 for quantitative variables, the working dataset has 9,277,579 records, and the corresponding file size is greater than 2GB. It is hard to apply any complicated model fitting with iterative algorithms on a single PC with limited memory.

Therefore, we turn to the developed divide-and-conquer strategy. It is natural to split the data by location, such as states or counties. According to our theoretical results, the number of sub-populations cannot be too large. The number of counties is more than 3,000 in U.S., while $\sqrt{9,277,579} \approx 3046$. Thus, we split the whole dataset by states and DC, resulting in 51 sub-populations. The number of records for each sub-population varies from 14,819 (Alaska) to 721,729 (California), and the median number is 128,247. It is reasonable to hypothesize that those categorical covariates are heterogeneous because their effects on the average amount that Medicare paid may vary across states. On the other hand, the outcome variable is the standardized payment by removing geographic differences in payment rates for individual services, and all three quan-

titative covariates are numbers of services and beneficiaries. Then it is reasonable to assume the effects of quantitative covariates are homogeneous.

We choose B-splines with degree of 3 to approximate the non-parametric functions of those three quantitative covariates. Assumption (A4) requires that the number of internal knots should be much smaller than $N^{\frac{1}{4}} \approx 55$. Additionally, we expect these curves are smooth. Thus, we set the number of internal knots as 5. Noting that the sizes of sub-populations are different, rather than a simple average to obtain the aggregated curves, a weighted average is employed by using weights $n_j / \sum_{j=1}^s n_j$, where n_j is the size of the j th sub-sample.

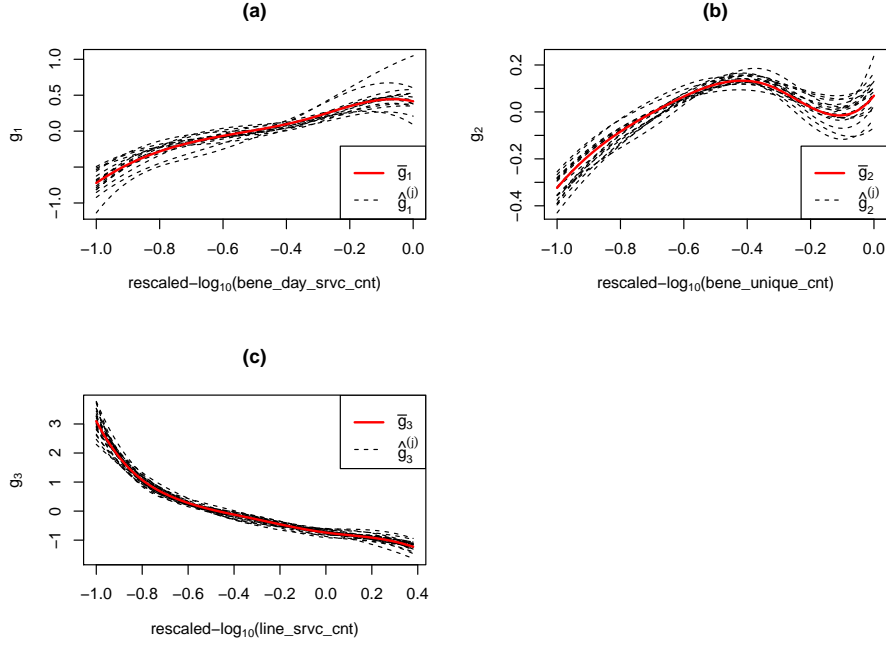
Figure 6: Box-plots of heterogeneous parameters across 51 states and DC: the left panel shows estimates of gender/entity, Medicare participation status, place of service and HCPCS drug status; the right panel shows estimates of 90 provider types versus the reference type.



Since the effects of those categorical covariates are allowed to be heterogeneous, we use box-plots to summarize the variabilities of their estimates across 51 sub-states. From Figure 6, which displays the extent of heterogeneity, we can see that only the effect of male versus female has small degree of heterogeneity around 0, and all the other estimates have substantial variabilities. It implies that the effects of most categorical covariates on the average amount that Medicare paid vary a lot across states.

Figure 7 presents the non-parametric estimates of the effects of those three quantitative covari-

Figure 7: Estimates of smooth functions based on each sub-population and the aggregation. (a): the estimated curves for “bene_unique_cnt”; (b): the estimated curves for “line_srvc_cnt”; (c): the estimated curves for “line_srvc_cnt”.



ates. The largest value of each quantitative covariate is different across states, so we only plot aggregated curves on the common support. From panels (a)-(c) of the figure, for each covariate, we can see estimated curves from 51 sub-samples (dashed lines in black color) are almost parallel to each other within a narrow band, while the aggregated curve (solid line in red color) is right in the middle of those sub-sample specific curves. Therefore, homogeneity assumption for these three quantitative covaraites seems reasonable.

5 Summary

In this paper, we develop a framework for additive partially linear models for massive heterogeneous data, using the divide-and-conquer strategy. As summarized in Wang et al. (2015), the divide-and-conquer strategy is one of the three commonly used strategies for analyzing massive data, with the other two being the sub-sampling strategy and the sequential updating strategy. However, the sub-sampling and sequential updating strategies are only suitable for analyzing ho-

mogeneous massive data. We can combine the divide-and-conquer and sub-sampling strategies to analyze heterogeneous data, by dividing the data into homogeneous subgroups and then conducting sub-sampling within each subgroup. We combine the divide-and-conquer and sequential updating strategies to analyze heterogeneous data, by dividing the data into homogeneous subgroups and then conducting sequential updating within each subgroup.

The framework developed in this paper extends the partially linear framework proposed in Zhao et al. (2016). Their partially linear framework considers only one common feature, using the smoothing-splines technique to fit the non-parametric function based on the general reproducing kernel Hilbert space (RKHS) theory (Wahba, 1990). Although the smoothing-splines technique and the RKHS theory have been well developed in the framework of generalized additive models Hastie and Tibshirani (1990), we find it very hard to extend them to our goal of analyzing massive data with multiple common features. Instead, we adopt polynomial splines for modeling the non-parametric effects of multiple common features simultaneously across all sub-populations while exploring heterogeneity of each sub-population. The proposed methods can be implemented easily and perform well in both simulation studies and the real data application. Here is a brief summary on the conditions of J_N that ensure those good asymptotic behaviours showed in Section 2.

First of all, all the theoretical results need Assumption (A4): $N^{\frac{1}{4p}} \ll J_N \ll N^{\frac{1}{4}}$. Besides this, different theorem (or corollary) needs different an extra condition. Here is the list of those conditions:

- (a) $J_N \ll n^{1/2}$;
- (b) $J_N \gg n^{1/(2p)}$;
- (c) $J_N \gg N^{1/(2p)}$ and $n \gg N^{1/2}$;
- (d) $J_N \gg N^{1/(2p)}$ and $J_N \ll s^{1/2}$.

In Theorem 1, under Condition (a), we derive the bound for the mean-square error of each sub-population specific estimator $\hat{g}^{(j)}$, $j = 1, \dots, s$. In Theorem 1, under Condition (b), we derive

the asymptotic normality for each sub-population specific estimator $\hat{\beta}^{(j)}$, $j = 1, \dots, s$. In Theorem 2, under Condition (a), we derive the bound for the mean-square error of the aggregated estimator \bar{g} . In Corollary 1, under Condition (c) and for the massive homogeneous data, we derive the asymptotic normality for the aggregated estimator $\bar{\beta}$. In Theorem 3, under Condition (d), we derive the asymptotic normality for each sub-population specific efficiency-boosted estimator $\check{\beta}^{(j)}$. These conditions can be satisfied by carefully selecting the balance between n and s , with some guidance provided in Remarks 1-5.

Appendix

A.1 Technical lemmas for Section 2.2

Define the centered version of B-spline basis as

$$b_{m,k}^*(z_k) = b_{m,k}(z_k) - \frac{E[b_{m,k}]}{E[b_{1,k}]} b_{1,k}(z_k), \quad k = 1, \dots, K, m = -\varrho + 1, \dots, J_N,$$

and the standardized version of B-spline basis as

$$B_{m,k}(z_k) = \frac{b_{m,k}^*(z_k)}{\|b_{m,k}^*\|_{2k}}, \quad m = -\varrho + 1, \dots, J_N, k = 1, \dots, K.$$

Then the minimization problem (3) is equivalent to the following minimization problem:

$$(\hat{\beta}^{(j)}, \hat{\gamma}^{(j)}) = \underset{\beta \in \mathbb{R}^d, \gamma \in \mathbb{R}^{K(J_N + \varrho)}}{\operatorname{argmin}} \frac{1}{2} \sum_{i \in G_j} [Y_i - \mathbf{X}_i^T \beta - \mathbf{B}(\mathbf{Z}_i)^T \gamma]^2,$$

where $\mathbf{B}(\mathbf{z}) = \{B_{m,k}(z_k), m = -\varrho + 1, \dots, J_N, k = 1, \dots, K\}^T$. Here, to abuse the notation, we still use $\hat{\gamma}^{(j)}$. Then $\hat{g}^{(j)}(\mathbf{z}) = \hat{\gamma}^T \mathbf{B}(\mathbf{z})$ is a spline estimator of g_0 for the j th sub-population, and

the centered spline estimators of a component function is

$$\hat{g}_k^{(j)}(z_k) = \sum_{m=-\varrho+1}^{J_N} \hat{\gamma}_{m,k} B_{m,k}(Z_k) - \frac{1}{n} \sum_{i \in G_j} \sum_{m=-\varrho+1}^{J_N} \hat{\gamma}_{m,k} B_{m,k}(Z_{ik}).$$

In practice, basis $\{b_{m,k}, m = -\varrho + 1, \dots, J_N, k = 1, \dots, K\}$ is used for computational implementation, while $\{B_{m,k}\}$ is convenient for asymptotic analysis.

De Boor (1978) showed that for any function $f \in \mathcal{H}$ and $N \geq 1$, there exists a function $\tilde{f} \in \mathcal{S}_N$ such that $\|\tilde{f} - f\|_\infty \leq Ch_N^p$. Thus, for g_0 satisfying Assumption (A1), there exists a $\tilde{g}^{(j)}(\mathbf{z}) = \mathbf{B}^T(\mathbf{z})\tilde{\gamma}^j \in \mathcal{G}_n^{(j)}$ s.t. $\|\tilde{g}^{(j)} - g_0\|_\infty = O(h_N^p)$ and $\tilde{g}^{(j)}(\mathbf{z})$ is the best least-squares projection of $g_0(\mathbf{z})$ into the space $\mathcal{G}_n^{(j)}$, implying

$$\langle \tilde{g}^{(j)}(\mathbf{z}) - g_0(\mathbf{z}), \mathbf{B}(\mathbf{z}) \rangle_{jn} = 0, \quad j = 1, \dots, s. \quad (\text{A.1})$$

Define

$$\tilde{\boldsymbol{\beta}}^{(j)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2} \sum_{i \in G_j} [Y_i - \tilde{g}^{(j)}(\mathbf{Z}_i) - \mathbf{X}_i^T \boldsymbol{\beta}]^2,$$

and let $m_{0i}^{(j)} \equiv m_0^{(j)}(\mathbf{T}_i) = g_0(\mathbf{Z}_i) + \mathbf{X}_i^T \boldsymbol{\beta}_0^{(j)}$, $\tilde{m}_0^{(j)}(\mathbf{t}) = \tilde{g}^{(j)}(\mathbf{z}) + \mathbf{x}^T \boldsymbol{\beta}_0^{(j)}$, and $\tilde{m}_{0i}^{(j)} \equiv \tilde{m}_0^{(j)}(\mathbf{T}_i) = \tilde{g}^{(j)}(\mathbf{Z}_i) + \mathbf{X}_i^T \tilde{\boldsymbol{\beta}}^{(j)}$.

Additionally, let $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\beta} \end{pmatrix}$, $\hat{\boldsymbol{\theta}}^{(j)} = \begin{pmatrix} \hat{\boldsymbol{\gamma}}^{(j)} \\ \hat{\boldsymbol{\beta}}^{(j)} \end{pmatrix}$, $\tilde{\boldsymbol{\theta}}^{(j)} = \begin{pmatrix} \tilde{\boldsymbol{\gamma}}^{(j)} \\ \tilde{\boldsymbol{\beta}}^{(j)} \end{pmatrix}$, $\hat{l}_n^{(j)}(\boldsymbol{\theta}) = l_n^{(j)}(\boldsymbol{\gamma}, \boldsymbol{\beta})$, and

$$\tilde{m}_i^{(j)} \equiv \tilde{m}^{(j)}(\mathbf{T}_i) = \tilde{g}^{(j)} + \mathbf{X}_i^T \tilde{\boldsymbol{\beta}} = \mathbf{B}^T(\mathbf{Z}_i) \tilde{\boldsymbol{\gamma}}^{(j)} + \mathbf{X}_i^T \tilde{\boldsymbol{\beta}}^{(j)}.$$

Define

$$V_n^{(j)} \triangleq \frac{\partial^2 \hat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \frac{1}{n} \sum_{i \in G_j} \left\{ \begin{array}{cc} (\mathbf{B}(\mathbf{Z}_i))^{\otimes 2} & \mathbf{B}(\mathbf{Z}_i) \mathbf{X}_i^T \\ \mathbf{X}_i \mathbf{B}^T(\mathbf{Z}_i) & \mathbf{X}_i^{\otimes 2} \end{array} \right\}.$$

Lemma A.1 Under Assumptions (A1)-(A4), for each sub-population j ,

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1} \boldsymbol{\Sigma}_1 \mathbf{A}^{-1}),$$

where $\mathbf{A} = E(\mathbf{X}^{\otimes 2})$ and $\boldsymbol{\Sigma}_1 = E(\varepsilon^2 \mathbf{X}^{\otimes 2})$.

Proof. Let $\tilde{\boldsymbol{\delta}}^{(j)} = \sqrt{n}(\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)})$. Then $\tilde{\boldsymbol{\delta}}^{(j)}$ minimizes

$$\tilde{l}_n^{(j)}(\boldsymbol{\delta}) = \frac{1}{2} \sum_{i \in G_j} \left[\left(Y_i - \tilde{m}_{0i}^{(j)} - \frac{1}{\sqrt{n}} \mathbf{X}_i^T \boldsymbol{\delta} \right)^2 - (Y_i - \tilde{m}_{0i}^{(j)})^2 \right].$$

Let $\mathbf{A}_n^{(j)} = \frac{1}{n} \sum_{i \in G_j} \mathbf{X}_i^{\otimes 2}$. By taking derivatives with respect to $\boldsymbol{\delta}$, we obtain

$$\frac{\partial \tilde{l}_n^{(j)}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \mathbf{A}_n^{(j)} \boldsymbol{\delta} - \frac{1}{\sqrt{n}} \sum_{i \in G_j} (Y_i - \tilde{m}_{0i}^{(j)}) \mathbf{X}_i = \mathbf{0},$$

which implies

$$\tilde{\boldsymbol{\delta}}^{(j)} = \frac{1}{\sqrt{n}} (\mathbf{A}_n^{(j)})^{-1} \sum_{i \in G_j} \varepsilon_i \mathbf{X}_i + \frac{1}{\sqrt{n}} (\mathbf{A}_n^{(j)})^{-1} \sum_{i \in G_j} (g_0(\mathbf{Z}_i) - \tilde{g}^{(j)}(\mathbf{Z}_i)) \mathbf{X}_i.$$

With similar arguments with those of Lemma A.1 in Liu et al. (2011) and the fact $\|\tilde{g}^{(j)} - g_0\|_\infty = O(h_N^p)$, the lemma follows. ■

Lemma A.2 Under Assumptions (A1)-(A4), if $J_N \ll \frac{n}{(\log n)^2}$, there exists a constant C such that $\sup_j \|(V_n^{(j)})^{-1}\|_2 \leq C$, a.s.

Proof. For each sub-population j , Lemma A.2 in Liu et al. (2011) showed there exists a constant C such that $\|(V_n^{(j)})^{-1}\|_2 \leq C$, a.s., if

$$\sup_{f_1, f_2 \in \mathcal{G}_n^{(j)}} \left| \frac{\langle f_1, f_2 \rangle_{jn} - \langle f_1, f_2 \rangle}{\|f_1\|_2 \|f_2\|_2} \right| = O\left(\frac{\log n}{(nh_N)^{1/2}}\right) = o(1), \quad \text{a.s.,}$$

by Lemma A.8 in Wang and Yang (2007). Here constant C is taken to be large enough to en-

sure that the above result holds for all $j = 1, \dots, s$. The condition $J_N \ll \frac{n}{(\log n)^2}$ implies $O(\log n / (nh_N)^{1/2}) = o(1)$. Therefore, the lemma is proved. \blacksquare

Lemma A.3 *Under Assumptions (A1)-(A4), for each sub-population j , we have*

$$\|\hat{\boldsymbol{\theta}}^{(j)} - \tilde{\boldsymbol{\theta}}^{(j)}\| = O_P \left(J_N^{1/2} n^{-1/2} + h_N^p \right).$$

Proof. It follows that

$$\left. \frac{\partial \hat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(j)}} - \left. \frac{\partial \hat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}^{(j)}} = \left. \frac{\partial^2 \hat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}^{(j)}} (\hat{\boldsymbol{\theta}}^{(j)} - \tilde{\boldsymbol{\theta}}^{(j)}),$$

where $\bar{\boldsymbol{\theta}}^{(j)}$ is between $\hat{\boldsymbol{\theta}}^{(j)}$ and $\tilde{\boldsymbol{\theta}}^{(j)}$. Thus, we have

$$\hat{\boldsymbol{\theta}}^{(j)} - \tilde{\boldsymbol{\theta}}^{(j)} = - \left(\left. \frac{\partial^2 \hat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}^{(j)}} \right)^{-1} \left. \frac{\partial \hat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}^{(j)}}.$$

We can write

$$\begin{aligned} \left. \frac{1}{n} \frac{\partial \hat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}^{(j)}} &= -\frac{1}{n} \sum_{i \in G_j} (Y_i - m_{0i}^{(j)}) \begin{pmatrix} B(\mathbf{Z}_i) \\ \mathbf{X}_i \end{pmatrix} \\ &\quad + \frac{1}{n} \sum_{i \in G_j} (\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \begin{pmatrix} B(\mathbf{Z}_i) \\ \mathbf{X}_i \end{pmatrix} \\ &\quad + \frac{1}{n} \sum_{i \in G_j} \mathbf{X}_i^T (\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \begin{pmatrix} B(\mathbf{Z}_i) \\ \mathbf{X}_i \end{pmatrix}. \end{aligned}$$

First, by (A.1), we have $\sum_{i \in G_j} (\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \mathbf{B}(\mathbf{Z}_i) = \mathbf{0}$. With similar arguments with those of Lemma A.3 in Liu et al. (2011), we have

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i \in G_j} (Y_i - m_{0i}^{(j)}) \mathbf{B}(\mathbf{Z}_i) \right\| &= O_P \left(J_N^{1/2} n^{-1/2} \right), \\ \left\| \frac{1}{n} \sum_{i \in G_j} (Y_i - m_{0i}^{(j)}) \mathbf{X}_i \right\| &= O_P \left(n^{-1/2} \right), \end{aligned}$$

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i \in G_j} (\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \mathbf{X}_i \right\| &= O_P(h_N^p), \\
\left\| \frac{1}{n} \sum_{i \in G_j} \mathbf{X}_i^T (\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \mathbf{B}(\mathbf{Z}_i) \right\| &= o_P(J_N^{1/2} n^{-1/2}), \\
\left\| \frac{1}{n} \sum_{i \in G_j} \mathbf{X}_i^T (\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \mathbf{X}_i \right\| &= o_P(n^{-1/2}).
\end{aligned}$$

Therefore, by Lemma A.2, we have

$$\|\hat{\boldsymbol{\theta}}^{(j)} - \tilde{\boldsymbol{\theta}}^{(j)}\| \leq \|(V_n^{(j)})^{-1}\|_2 \left\| \frac{1}{n} \frac{\partial \hat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}^{(j)}} \right\| = O_P(J_N^{1/2} n^{-1/2} + h_N^p). \blacksquare$$

Lemma A.4 Under Assumptions (A1)-(A4), for each sub-population j , if $J_N \ll \frac{n}{(\log n)^2}$, we have

$$\begin{aligned}
\|\hat{g}^{(j)} - g_0\|_2 &= O_P(J_N^{1/2} n^{-1/2} + h_N^p), \\
\|\hat{g}^{(j)} - g_0\|_{jn} &= O_P(J_N^{1/2} n^{-1/2} + h_N^p), \\
\|\hat{g}_k^{(j)} - g_{0k}\|_{2k} &= O_P(J_N^{1/2} n^{-1/2} + h_N^p), \\
\|\hat{g}_k^{(j)} - g_{0k}\|_{jnk} &= O_P(J_N^{1/2} n^{-1/2} + h_N^p),
\end{aligned}$$

where $k = 1, \dots, K$.

Proof. The proof is similar with that of Lemma A.4 in Liu et al. (2011) by applying Lemmas A.2 and A.3 and noting that

$$\sup_{f \in \mathcal{S}_N} \frac{\|f\|_{jnk}}{\|f\|_{2k}} = O_P\left(\frac{\log n}{(nh_N)^{1/2}}\right) = o_P(1), \quad k = 1, \dots, K,$$

which is implied by Lemma A.8 in Wang and Yang (2007) under condition $J_N \ll n/(\log n)^2$. \blacksquare

Lemma A.5 Under Assumptions (A1)-(A4), for each sub-population j , if $n \gg J_N^2$, we have

$$\frac{1}{n} \sum_{i \in G_j} \widetilde{\mathbf{X}}_i \Gamma(\mathbf{Z}_i)^T (\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) = o_P(n^{-1/2}),$$

$$\frac{1}{n} \sum_{i \in G_j} (\hat{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \widetilde{\mathbf{X}}_i = o_P(n^{-1/2}),$$

Proof. The proof is similar with that of Lemma A.5 in Liu et al. (2011) by making following revisions. We only show the second equality, and the first one can be proved similarly.

Let $w_1(\mathbf{Z}, g) = g(\mathbf{Z})\widetilde{\mathbf{X}}$, and it follows

$$E\|w_1(\mathbf{Z}, \hat{g}^{(j)}) - w_1(\mathbf{Z}, g_0)\|^2 = E\|(\hat{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i))\widetilde{\mathbf{X}}_i\|^2 \leq O(E\|\hat{g}^{(j)} - g_0\|_2^2).$$

By Lemma A.2 of Huang (1999), the logarithm of the ε -bracketing number of the class of functions $\mathcal{A}_1^{(j)}(\delta) = \{w_1(\cdot, \hat{g}) - w_1(\cdot, g_0) : \hat{g} \in \mathcal{G}_n^{(j)}, \|\hat{g} - g_0\|_2 \leq \delta\}$ is $c\{(J_N - \varrho)\log(\delta/\varepsilon) + \log(\delta^{-1})\}$. Thus, the corresponding entropy integral $J_{[]}(\delta, \mathcal{A}_1^{(j)}(\delta), \|\cdot\|_2) \leq c\delta\{(J_N - \varrho)^{1/2} + \log^{1/2}(\delta^{-1})\}$. According to Lemma 7 of Stone (1986) and Lemma A.4, $\|\hat{g}^{(j)} - g_0\|_\infty \leq cJ_N^{1/2}\|\hat{g}^{(j)} - g_0\|_2 = O_P(J_N n^{-1/2} + J_N^{1/2} h_N^p)$. Let $r_{n,N}^{-1} = J_N^{1/2} n^{-1/2} + h_N^p$, then

$$\begin{aligned} & E \left| \frac{1}{n} \sum_{i \in G_j} \{\hat{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)\} \widetilde{\mathbf{X}}_i - E \{\hat{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)\} \widetilde{\mathbf{X}}_i \right| \\ & \leq n^{-1/2} C r_{n,N}^{-1} \left\{ (J_N + \varrho)^{1/2} + \log^{1/2}(r_{n,N}) \right\} \\ & \quad \times \left[1 + \frac{c r_{n,N}^{-1} \left\{ (J_N + \varrho)^{1/2} + \log^{1/2}(r_{n,N}) \right\}}{r_{n,N}^{-2} \sqrt{n}} C_0 \right] \\ & \leq O(1) n^{-1/2} C r_{n,N}^{-1} \left\{ (J_N + \varrho)^{1/2} + \log^{1/2}(r_{n,N}) \right\}, \end{aligned}$$

where the second inequality is based on the fact $r_{n,N} J_N^{1/2} / \sqrt{n} = O(1)$.

Under condition that $n \gg J_N^2$, we have $J_N / \sqrt{n} = o(1)$, implying $J_N n^{-1/2} + J_N^{1/2} h_N^p = o(1)$, and then $r_{n,N}^{-1} J_N^{1/2} = o(1)$. Therefore, the above term is bounder by $o(n^{-1/2})$. \blacksquare

A.2 Technical lemmas for Section 2.3

Let $\tilde{g} = \frac{1}{s} \sum_{j=1}^s \tilde{g}^{(j)}$. In order to ensure that $\tilde{g} \in \mathcal{G}_N$, we re-center the individual estimator $\tilde{g}^{(j)}(\mathbf{z})$ via $\tilde{g}^{(j)}(\mathbf{z}) = \sum_{k=1}^K \sum_{m=-\varrho}^{J_N} \tilde{\gamma}_{m,k} b_{m,k}(z_k) - \frac{1}{N} \sum_{k=1}^K \sum_{i=1}^N \sum_{m=-\varrho}^{J_N} \tilde{\gamma}_{m,k} b_{m,k}(z_{ik})$. To abuse the no-

tation, we still denote the centered estimator as $\tilde{g}^{(j)}(\mathbf{z})$. Lemma A.2 shows $\|(V_n^{(j)})^{-1}\|_2 \leq C$ a.s., $j = 1, \dots, s$, if $n \gg J_N^2$. This property plays a key role in all following proofs. Define $\mathbf{u}_i^{(j)} = (V_n^{(j)})^{-1} \begin{pmatrix} \mathbf{B}(\mathbf{Z}_i) \\ \mathbf{X}_i \end{pmatrix} = \{u_{im}^{(j)}, m = 1, \dots, K(J_N + \varrho) + d\}^T$, and it follows $\sum_{i \in G_j} (\mathbf{u}_i^{(j)})^{\otimes 2} = n(V_n^{(j)})^{-1}$. Let \mathbf{e}_m denote a $(K(J_N + \varrho) + d)$ -dim vector with its m th entry as 1 and 0 otherwise, and thus $u_{im}^{(j)} = \mathbf{e}_m^T \mathbf{u}_i^{(j)}$.

Lemma A.6 *Under Assumptions (A1)-(A4), if $n \gg J_N^2$, we have*

$$\left\| \frac{1}{N} \sum_{j=1}^s \sum_{i \in G_j} \varepsilon_i \mathbf{u}_i^{(j)} \right\| = O_P(J_N^{1/2} N^{-1/2}).$$

Proof. It follows

$$\left\| \frac{1}{N} \sum_{j=1}^s \sum_{i \in G_j} \varepsilon_i \mathbf{u}_i^{(j)} \right\|^2 = \frac{1}{N^2} \sum_{m=1}^{K(J_N + \varrho) + d} \left\{ \sum_{j=1}^s \sum_{i \in G_j} \varepsilon_i u_{im}^{(j)} \right\}^2.$$

Observing that

$$\begin{aligned} \frac{1}{N} E \left\{ \sum_{j=1}^s \sum_{i \in G_j} \varepsilon_i u_{im}^{(j)} \right\}^2 &= \frac{1}{N^2} E \left[\sum_{j=1}^s \sum_{i \in G_j} \varepsilon_i^2 (u_{im}^{(j)})^2 \right] = \frac{\sigma^2}{N^2} E \left[\sum_{j=1}^s \sum_{i \in G_j} (u_{im}^{(j)})^2 \right] \\ &= \frac{\sigma^2}{N^2} E \left[\sum_{j=1}^s n \mathbf{e}_m^T (V_n^{(j)})^{-1} \mathbf{e}_m \right] \leq \frac{NC\sigma^2}{N^2}, \end{aligned}$$

where the last inequality is due to the fact $\mathbf{e}_m^T (V_n^{(j)})^{-1} \mathbf{e}_m \leq \|(V_n^{(j)})^{-1}\|_2^2 \|\mathbf{e}_m\|^2 \leq C$ a.s.. Thus

$$\left\| \frac{1}{N} \sum_{j=1}^s \sum_{i \in G_j} \varepsilon_i \mathbf{u}_i^{(j)} \right\| = O_P(J_N^{1/2} N^{-1/2}). \blacksquare$$

Lemma A.7 *Under Assumptions (A1)-(A4), if $n \gg J_N^2$, we have*

$$\left\| \frac{1}{N} \sum_{j=1}^s \sum_{i \in G_j} (\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \mathbf{u}_i^{(j)} \right\| = O_P(J_N^{1/2} h_N^p).$$

Proof. Note that

$$\begin{aligned} \frac{1}{N} \sum_{j=1}^s \sum_{i \in G_j} (\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \mathbf{u}_i^{(j)} &= \frac{1}{N} \sum_{j=1}^s (V_n^{(j)})^{-1} \sum_{i \in G_j} (\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \begin{pmatrix} \mathbf{B}(\mathbf{Z}_i) \\ \mathbf{X}_i \end{pmatrix} \\ &= \frac{1}{N} \sum_{j=1}^s (V_n^{(j)})^{-1} \sum_{i \in G_j} (\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \begin{pmatrix} \mathbf{0} \\ \mathbf{X}_i \end{pmatrix}, \end{aligned}$$

where the last equality follows from (A.1).

Therefore, it follows

$$\begin{aligned} \left\| \frac{1}{N} \sum_{j=1}^s \sum_{i \in G_j} (\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \mathbf{u}_i^{(j)} \right\| &\leq \frac{1}{N} \sum_{j=1}^s \|(V_n^{(j)})^{-1}\|_2 \sum_{i \in G_j} \|(\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \begin{pmatrix} \mathbf{0} \\ \mathbf{X}_i \end{pmatrix}\| \\ &\leq \frac{C}{N} \sum_{j=1}^s \sum_{i \in G_j} \|(\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i))\|_\infty \left\| \begin{pmatrix} \mathbf{0} \\ \mathbf{X}_i \end{pmatrix} \right\|_1 = O_P(h_N^p). \blacksquare \end{aligned}$$

Lemma A.8 Under Assumptions (A1)-(A4), if $n \gg J_N^2$, we have

$$\left\| \frac{1}{N} \sum_{j=1}^s \sum_{i \in G_j} \mathbf{X}_i^T (\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \mathbf{u}_i^{(j)} \right\| = o_P \left(J_N^{1/2} N^{-1/2} \right).$$

Proof. It follows

$$\left\| \frac{1}{N} \sum_{j=1}^s \sum_{i \in G_j} \mathbf{X}_i^T (\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \mathbf{u}_i^{(j)} \right\|^2 = \frac{1}{N^2} \sum_{m=1}^{K(J_N + \varrho) + d} \left\{ \sum_{j=1}^s \sum_{i \in G_j} \mathbf{X}_i^T (\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) u_{im}^{(j)} \right\}^2.$$

The proof of Lemma A.1 shows

$$\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)} = \frac{1}{n} (\mathbf{A}_n^{(j)})^{-1} \sum_{i \in G_j} \varepsilon_i \mathbf{X}_i + \frac{1}{n} (\mathbf{A}_n^{(j)})^{-1} \sum_{i \in G_j} (\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \mathbf{X}_i.$$

Then

$$\sum_{i \in G_j} \mathbf{X}_i^T (\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) u_{im}^{(j)}$$

$$\begin{aligned}
&= \sum_{i_1 \in G_j} u_{i_1 m}^{(j)} \mathbf{X}_{i_1}^T \left(\frac{1}{n} (\mathbf{A}_n^{(j)})^{-1} \sum_{i_2 \in G_j} \varepsilon_{i_2} \mathbf{X}_{i_2} + \frac{1}{n} (\mathbf{A}_n^{(j)})^{-1} \sum_{i_2 \in G_j} (\tilde{g}^{(j)}(\mathbf{Z}_{i_2}) - g_0(\mathbf{Z}_{i_2})) \mathbf{X}_{i_2} \right) \\
&= \frac{1}{n} \sum_{i_2 \in G_j} \varepsilon_{i_2} \mathbf{X}_{i_2}^T \left(\sum_{i_1 \in G_j} u_{i_1 m}^{(j)} (\mathbf{A}_n^{(j)})^{-1} \mathbf{X}_{i_1} \right) \\
&\quad + \frac{1}{n} \sum_{i_2 \in G_j} (\tilde{g}^{(j)}(\mathbf{Z}_{i_2}) - g_0(\mathbf{Z}_{i_2})) \mathbf{X}_{i_2}^T \left(\sum_{i_1 \in G_j} u_{i_1 m}^{(j)} (\mathbf{A}_n^{(j)})^{-1} \mathbf{X}_{i_1} \right) \\
&= \sum_{i \in G_j} \varepsilon_i \mathbf{X}_i^T \mathbf{v}_m^{(j)} + \sum_{i \in G_j} (\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \mathbf{X}_i^T \mathbf{v}_m^{(j)},
\end{aligned}$$

where $\mathbf{v}_m^{(j)} = \sum_{i \in G_j} u_{im}^{(j)} (\mathbf{A}_n^{(j)})^{-1} \mathbf{X}_i$. Thus, we have

$$\begin{aligned}
&\frac{1}{N^2} E \left\{ \sum_{j=1}^s \sum_{i \in G_j} \mathbf{X}_i^T (\tilde{\beta}^{(j)} - \beta_0^{(j)}) u_{im}^{(j)} \right\}^2 \\
&= \frac{1}{N^2} E \left\{ \sum_{j=1}^s \sum_{i \in G_j} \varepsilon_i \mathbf{X}_i^T \mathbf{v}_m^{(j)} + \sum_{j=1}^s \sum_{i \in G_j} (\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \mathbf{X}_i^T \mathbf{v}_m^{(j)} \right\}^2 \\
&= \frac{1}{N^2} E \left\{ \sum_{j=1}^s \sum_{i \in G_j} \varepsilon_i \mathbf{X}_i^T \mathbf{v}_m^{(j)} \right\}^2 + \frac{1}{N^2} E \left\{ \sum_{j=1}^s \sum_{i \in G_j} (\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \mathbf{X}_i^T \mathbf{v}_m^{(j)} \right\}^2 \\
&\leq \frac{\sigma^2}{N^2} E \sum_{j=1}^s \sum_{i \in G_j} \{ \mathbf{X}_i^T \mathbf{v}_m^{(j)} \}^2 + \frac{1}{N} \|\tilde{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)\|_\infty^2 E \sum_{j=1}^s \sum_{i \in G_j} \{ \mathbf{X}_i^T \mathbf{v}_m^{(j)} \}^2.
\end{aligned}$$

Observing that

$$\begin{aligned}
\sum_{i \in G_j} \{ \mathbf{X}_i^T \mathbf{v}_m^{(j)} \}^2 &= \sum_{i \in G_j} (\mathbf{v}_m^{(j)})^T \mathbf{X}_i \mathbf{X}_i^T \mathbf{v}_m^{(j)} = n (\mathbf{v}_m^{(j)})^T \mathbf{A}_n^{(j)} \mathbf{v}_m^{(j)} \\
&= \frac{1}{n} \left(\sum_{i \in G_j} u_{im}^{(j)} \mathbf{X}_i^T (\mathbf{A}_n^{(j)})^{-1} \right) \mathbf{A}_n^{(j)} \left(\sum_{i \in G_j} u_{im}^{(j)} \mathbf{A}_n^{(j)} \mathbf{X}_i \right) \\
&= \left(\frac{1}{\sqrt{n}} \sum_{i \in G_j} u_{im}^{(j)} \mathbf{X}_i \right)^T (\mathbf{A}_n^{(j)})^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i \in G_j} u_{im}^{(j)} \mathbf{X}_i \right)
\end{aligned}$$

Based on the Central Limit Theorem and Slutsky's Theorem, it follows

$$E \left\{ \sum_{i \in G_j} (\mathbf{X}_i^T \mathbf{v}_m^{(j)})^2 \right\} = O(1).$$

Therefore,

$$\frac{1}{N^2} E \left\{ \sum_{j=1}^s \sum_{i \in G_j} \mathbf{X}_i^T (\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \mathbf{u}_{im}^{(j)} \right\}^2 = O \left(\frac{s}{N^2} + \frac{sh_N^{2p}}{N} \right) = o \left(\frac{1}{N} \right),$$

noting that $h_N^{2p} \ll N^{-1}$. Then, we have

$$\left\| \frac{1}{N} \sum_{j=1}^s \sum_{i \in G_j} \mathbf{X}_i^T (\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \mathbf{u}_i^{(j)} \right\| = o_P \left(J_N^{1/2} N^{-1/2} \right). \blacksquare$$

A.3 Proofs of theorems

Proof of Theorem 1. The results about $\hat{g}^{(j)}$ are implied by Lemma A.4 directly. We only need to prove the stated result about $\hat{\boldsymbol{\beta}}^{(j)}$. Note that the condition that $J_N^2 \ll n$ implies that $J_N \ll \frac{n}{(\log n)^2}$. Also the condition that $J_N \gg n^{1/(2p)}$ implies that $h_N^p = O(J_N^{-p}) \ll n^{-1/2}$. Therefore, the stated result about $\hat{\boldsymbol{\beta}}^{(j)}$ can be showed by Lemmas A.1-A.5, following the same argument of proving Theorem 1 in Liu et al. (2011). \blacksquare

Proof of Theorem 2. We first quantify $\|\hat{g}^{(j)} - g_0\|_2$. Noting $\|\tilde{g} - g_0\|_2 \leq \|\tilde{g} - g_0\|_\infty = O(h_N^p)$ and

$$\frac{1}{s} \sum_{j=1}^s (\hat{g}^{(j)}(z) - \tilde{g}^{(j)}(z)) = \frac{1}{s} \mathbf{B}^T(z) \sum_{j=1}^s (\hat{\boldsymbol{\gamma}}^{(j)} - \tilde{\boldsymbol{\gamma}}^{(j)}),$$

we have

$$\begin{aligned} \left\| \frac{1}{s} \sum_{j=1}^s (\hat{g}^{(j)}(z) - \tilde{g}^{(j)}(z)) \right\|_2^2 &= \int_{[0,1]^K} \left[\frac{1}{s} \sum_{j=1}^s (\hat{g}^{(j)}(z) - \tilde{g}^{(j)}(z)) \right]^2 f(z) dz \\ &= \frac{1}{s} \sum_{j=1}^s (\hat{\boldsymbol{\gamma}}^{(j)} - \tilde{\boldsymbol{\gamma}}^{(j)})^T [E \mathbf{B}(z) \mathbf{B}^T(z)] \sum_{j=1}^s (\hat{\boldsymbol{\gamma}}^{(j)} - \tilde{\boldsymbol{\gamma}}^{(j)}) \end{aligned}$$

$$\leq \frac{C}{s^2} \left\| \sum_{j=1}^s (\hat{\gamma}^{(j)} - \tilde{\gamma}^{(j)}) \right\|^2 \leq \frac{C}{s^2} \left\| \sum_{j=1}^s (\hat{\boldsymbol{\theta}}^{(j)} - \tilde{\boldsymbol{\theta}}^{(j)}) \right\|^2.$$

Then we consider $\left\| \frac{1}{s} \sum_{j=1}^s (\hat{\boldsymbol{\theta}}^{(j)} - \tilde{\boldsymbol{\theta}}^{(j)}) \right\|$. The proof of Lemma A.3 implies that

$$\hat{\boldsymbol{\theta}}^{(j)} - \tilde{\boldsymbol{\theta}}^{(j)} = (V_n^{(j)})^{-1} \frac{1}{n} \frac{\partial \hat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}^{(j)}},$$

where $\frac{1}{n} \frac{\partial \hat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}^{(j)}}$ is equal to

$$\left\{ -\frac{1}{n} \sum_{i \in G_j} (Y_i - m_{0i}^{(j)}) + \frac{1}{n} \sum_{i \in G_j} (\hat{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) + \frac{1}{n} \sum_{i \in G_j} \mathbf{X}_i^T (\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \right\} \begin{pmatrix} \mathbf{B}(\mathbf{Z}_i) \\ \mathbf{X}_i \end{pmatrix}.$$

Then $\frac{1}{s} \sum_{j=1}^s (\hat{\boldsymbol{\theta}}^{(j)} - \tilde{\boldsymbol{\theta}}^{(j)})$ is equal to

$$-\frac{1}{N} \sum_{j=1}^s \sum_{i \in G_j} \varepsilon_i \mathbf{u}_i^{(j)} + \frac{1}{N} \sum_{j=1}^s \sum_{i \in G_j} (\hat{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \mathbf{u}_i^{(j)} + \frac{1}{N} \sum_{j=1}^s \sum_{i \in G_j} \mathbf{X}_i^T (\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \mathbf{u}_i^{(j)}.$$

Therefore, combining Lemmas A.6-A.8, we have

$$\frac{1}{s} \left\| \sum_{j=1}^s (\hat{\boldsymbol{\theta}}^{(j)} - \tilde{\boldsymbol{\theta}}^{(j)}) \right\| = O_P \left(J_N^{1/2} N^{-1/2} + h_N^p \right).$$

and further we have,

$$\begin{aligned} \|\bar{g} - g_0\|_2 &= \left\| \frac{1}{s} \sum_{j=1}^s \hat{g}^{(j)} - \frac{1}{s} \sum_{j=1}^s \tilde{g}^{(j)} + \tilde{g} - g_0 \right\|_2 \leq \left\| \frac{1}{s} \sum_{j=1}^s (\hat{g}^{(j)} - \tilde{g}^{(j)}) \right\|_2 + \|\tilde{g} - g_0\|_2 \\ &= O_P \left(J_N^{1/2} N^{-1/2} + h_N^p \right). \end{aligned}$$

Next we quantify $\|\bar{g} - g_0\|_N$. Using Lemma A.8 in Wang and Yang (2007), we have

$$C_N \equiv \sup_{f_1, f_2 \in \mathcal{G}_N} \left| \frac{\langle f_1, f_2 \rangle_N - \langle f_1, f_2 \rangle}{\|f_1\| \|f_2\|} \right| = O \left(\frac{\log N}{(N h_N)^{1/2}} \right), \quad a.s.$$

Therefore, noting $\bar{g}, \tilde{g} \in \mathcal{G}_N$, we have

$$\|\bar{g} - g_0\|_N \leq \|\bar{g} - \tilde{g}\|_N + \|\tilde{g} - g_0\|_N = O_P\left(J_N^{1/2}N^{-1/2} + h_N^p\right). \quad \blacksquare$$

Proof of Corollary 1. For homogenous massive data, $\hat{\beta}^{(j)}$, $j = 1, \dots, s$, are i.i.d. random vectors. Li et al. (2013) showed that if $E(\hat{\beta}^{(j)} - \beta_0) = o(N^{-1/2})$, then $\bar{\beta}$ is as efficient as $\hat{\beta}$, which is defined as via the following minimization using all N observations,

$$(\hat{\beta}, \hat{g}) = \underset{\beta \in \mathbb{R}^d, g \in \mathcal{G}_N}{\operatorname{argmin}} \frac{1}{2} \sum_{j=1}^s \sum_{i \in G_j} [Y_i - \mathbf{X}_i^T \beta - g(\mathbf{Z}_i)]^2.$$

Liu et al. (2011) showed that

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{D}^{-1}).$$

Therefore it suffices to show that $E(\hat{\beta}^{(j)} - \beta_0) = o(N^{-1/2})$. Following the proof of Theorem 1, we have

$$\begin{aligned} & \left(\frac{1}{n} \sum_{i \in G_j} \tilde{\mathbf{X}}_i^{\otimes 2} + \frac{1}{n} \sum_{i \in G_j} \tilde{\mathbf{X}}_i \Gamma(\mathbf{Z}_i)^T \right) (\hat{\beta}^{(j)} - \beta_0) \\ &= \frac{1}{n} \sum_{i \in G_j} \varepsilon_i \tilde{\mathbf{X}}_i - \frac{1}{n} \sum_{i \in G_j} (\hat{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \tilde{\mathbf{X}}_i, \end{aligned}$$

and it follows

$$\begin{aligned} \hat{\beta}^{(j)} - \beta_0 &= \left(\frac{1}{n} \sum_{i \in G_j} \tilde{\mathbf{X}}_i^{\otimes 2} + \frac{1}{n} \sum_{i \in G_j} \tilde{\mathbf{X}}_i \Gamma(\mathbf{Z}_i)^T \right)^{-1} \times \\ & \quad \left[\frac{1}{n} \sum_{i \in G_j} \varepsilon_i \tilde{\mathbf{X}}_i - \frac{1}{n} \sum_{i \in G_j} (\hat{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \tilde{\mathbf{X}}_i \right]. \end{aligned}$$

Under Assumption A3 and the fact that $E(\phi(\mathbf{Z})\tilde{\mathbf{X}}) = 0$ for any measurable function ϕ , we can

show that

$$0 < c \leq E \left\| \frac{1}{n} \sum_{i \in G_j} \widetilde{\mathbf{X}}_i^{\otimes 2} + \frac{1}{n} \sum_{i \in G_j} \widetilde{\mathbf{X}}_i \Gamma(\mathbf{Z}_i)^T \right\|_2 \leq C,$$

where c and C are some positive constants. Moreover, we have

$$E \left\{ \frac{1}{n} \sum_{i \in G_j} (\widehat{g}^{(j)}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \widetilde{\mathbf{X}}_i \right\} = E (\widehat{g}^{(j)}(\mathbf{Z}) - g_0(\mathbf{Z})) \widetilde{\mathbf{X}}_i = O(h_N^p) = o(N^{-1/2}).$$

Therefore, if $n \gg N^{1/2}$, by Cauchy-Schwarz inequality we have $E(\widehat{\beta}^{(j)} - \beta_0) = o(N^{-1/2})$. \blacksquare

Proof of Theorem 3. The estimating equation is

$$\sum_{i \in G_j} \mathbf{X}_i \left\{ Y_i - \mathbf{X}_i^T \check{\beta}^{(j)} - \bar{g}(\mathbf{Z}_i) \right\} = \mathbf{0}.$$

Since $Y_i = \mathbf{X}_i^T \beta_0^{(j)} + g_0(\mathbf{Z}_i) + \varepsilon_i$, we have

$$\sqrt{n} (\check{\beta}^{(j)} - \beta_0^{(j)}) = n^{-1/2} \sum_{i \in G_j} (\mathbf{A}_n^{(j)})^{-1} \mathbf{X}_i^T \varepsilon_i + n^{-1/2} \sum_{i \in G_j} (\mathbf{A}_n^{(j)})^{-1} \mathbf{X}_i (g_0(\mathbf{Z}_i) - \bar{g}(\mathbf{Z}_i)).$$

Considering the first term on the right hand side of the above equation, we have

$$n^{-1/2} \sum_{i \in G_j} (\mathbf{A}_n^{(j)})^{-1} \mathbf{X}_i^T \varepsilon_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{A}^{-1}).$$

Consider the second term on the right hand side. Let $w_2(\mathbf{Z}, g) = g(\mathbf{Z}) (\mathbf{A}_n^{(j)})^{-1} \mathbf{X}$. We have

$$E \|w_2(\mathbf{Z}, \bar{g}) - w_2(\mathbf{Z}, g_0)\|^2 = E \left\| (\bar{g}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) (\mathbf{A}_n^{(j)})^{-1} \mathbf{X}_i \right\|^2 \leq O(E \|\bar{g} - g_0\|_2^2).$$

By Lemma A.2 of Huang (1999), the logarithm of the ε -bracketing number of the class of functions $\mathcal{A}_2(\delta) = \{w_2(\cdot, \bar{g}) - s(\cdot, g_0) : \bar{g} \in \mathcal{G}_N, \|\bar{g} - g_0\|_2 \leq \delta\}$ is $c\{(J_N - \varrho)\log(\delta/\varepsilon) + \log(\delta^{-1})\}$. Thus, the corresponding entropy integral $J_{[\cdot]}(\delta, \mathcal{A}_2(\delta), \|\cdot\|_2) \leq c\delta\{(J_N - \varrho)^{1/2} + \log^{1/2}(\delta^{-1})\}$. According to Lemma 7 of Stone (1986) and Theorem 2, $\|\bar{g} - g_0\|_\infty \leq cJ_N^{1/2} \|\bar{g} - g_0\|_2 = O_P(J_N N^{-1/2} + J_N^{1/2} h_N^p)$.

Let $r_N^{-1} = J_N^{1/2} N^{-1/2} + h_N^p$, then

$$\begin{aligned}
& E \left| \frac{1}{n} \sum_{i \in G_j} \left\{ (\bar{g}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \left(\mathbf{A}_n^{(j)} \right)^{-1} \mathbf{X}_i \right\} \right. \\
& \quad \left. - E \left\{ (\bar{g}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \left(\mathbf{A}_n^{(j)} \right)^{-1} \mathbf{X}_i \right\} \right| \\
& \leq n^{-1/2} C r_N^{-1} \left\{ (J_N + \varrho)^{1/2} + \log^{1/2}(r_N) \right\} \\
& \quad \times \left[1 + \frac{c r_N^{-1} \left\{ (J_N + \varrho)^{1/2} + \log^{1/2}(r_N) \right\}}{r_N^{-2} \sqrt{n}} C_0 \right] \\
& \leq O(s^{1/2}) n^{-1/2} C r_N^{-1} \left\{ (J_N + \varrho)^{1/2} + \log^{1/2}(r_N) \right\} \\
& = O(n^{-1/2}) \times O \left(J_N n^{-1/2} + s^{1/2} J_N^{1/2} h_N^p \right) \\
& = O(n^{-1/2}) \times O \left(J_N n^{-1/2} + (n^{-1} N^{1+q(1-2p)})^{1/2} \right) \\
& \leq O(n^{-1/2}) \times O \left(J_N n^{-1/2} + (n^{-1} N^{1/(2p)})^{1/2} \right),
\end{aligned}$$

where the last inequality is due to the condition that $J_N \gg N^{1/(2p)}$. The condition that $J_N^2 \ll n$ implies that $O(J_N n^{-1/2}) = o(1)$ and $n \gg N^{1/(2p)}$ to make sure that the above expectation has an order $o(n^{-1/2})$. Furthermore,

$$E \left\{ (\bar{g}(\mathbf{Z}_i) - g_0(\mathbf{Z}_i)) \left(\mathbf{A}_n^{(j)} \right)^{-1} \mathbf{X}_i \right\} \leq O(E \|\bar{g} - g_0\|_\infty) = O(J_N N^{-1/2}).$$

Thus,

$$n^{-1/2} \sum_{i \in G_j} \left(\mathbf{A}_n^{(j)} \right)^{-1} \mathbf{X}_i (g_0(\mathbf{Z}_i) - \bar{g}(\mathbf{Z}_i)) = O(J_N s^{-1/2}) + o_P(1) = o_P(1),$$

where the last equality is due to the condition that $J_N^2 \ll s$. Therefore, the theorem is proved. \blacksquare

Proof of Theorem 4. Under the null hypothesis, we have

$$\sqrt{n} \mathbf{Q}(\hat{\boldsymbol{\beta}}^{(j_1)} - \hat{\boldsymbol{\beta}}^{(j_2)}) = \sqrt{n} \mathbf{Q}(\hat{\boldsymbol{\beta}}^{(j_1)} - \boldsymbol{\beta}_0^{(j_1)}) - \sqrt{n} \mathbf{Q}(\hat{\boldsymbol{\beta}}^{(j_2)} - \boldsymbol{\beta}_0^{(j_2)}).$$

By Theorem 1, we have

$$\sqrt{n}\mathbf{Q}(\hat{\beta}^{(jt)} - \beta_0^{(jt)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{Q} \mathbf{D}^{-1} \mathbf{Q}^T),$$

where $t = 1$ or 2 . Therefore, $\sqrt{n}\mathbf{Q}(\hat{\beta}^{(j_1)} - \hat{\beta}^{(j_2)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, 2\sigma^2 \mathbf{Q} \mathbf{D}^{-1} \mathbf{Q}^T)$.

Consider the second result. According to the proof of Theorem 3, we have

$$\begin{aligned} \sqrt{n}(\check{\beta}^{(j)} - \beta_0^{(j)}) &= n^{-1/2} \sum_{i \in G_j} \left(\mathbf{A}_n^{(j)} \right)^{-1} \mathbf{X}_i^T \varepsilon_i \\ &\quad + n^{-1/2} \sum_{i \in G_j} \left(\mathbf{A}_n^{(j)} \right)^{-1} \mathbf{X}_i (g_0(\mathbf{Z}_i) - \bar{g}(\mathbf{Z}_i)). \end{aligned}$$

Thus, with similar arguments in the proof of Theorem 3, we have

$$\begin{aligned} \sqrt{n}\mathbf{Q}(\check{\beta}^{(j_1)} - \check{\beta}^{(j_2)}) &= \sqrt{n}\mathbf{Q}(\check{\beta}^{(j_1)} - \beta_0^{(j_1)}) - \sqrt{n}\mathbf{Q}(\check{\beta}^{(j_2)} - \beta_0^{(j_2)}) \\ &= n^{-1/2} \sum_{i \in G_{j_1}} \mathbf{Q} \left(\mathbf{A}_n^{(j_1)} \right)^{-1} \mathbf{X}_i^T \varepsilon_i + n^{-1/2} \sum_{i \in G_{j_1}} \mathbf{Q} \left(\mathbf{A}_n^{(j_1)} \right)^{-1} \mathbf{X}_i (g_0(\mathbf{Z}_i) - \bar{g}(\mathbf{Z}_i)) \\ &\quad - n^{-1/2} \sum_{i \in G_{j_2}} \mathbf{Q} \left(\mathbf{A}_n^{(j_2)} \right)^{-1} \mathbf{X}_i^T \varepsilon_i - n^{-1/2} \sum_{i \in G_{j_2}} \mathbf{Q} \left(\mathbf{A}_n^{(j_2)} \right)^{-1} \mathbf{X}_i (g_0(\mathbf{Z}_i) - \bar{g}(\mathbf{Z}_i)) \\ &= n^{-1/2} \sum_{i \in G_{j_1}} \mathbf{Q} \left(\mathbf{A}_n^{(j_1)} \right)^{-1} \mathbf{X}_i^T \varepsilon_i - n^{-1/2} \sum_{i \in G_{j_2}} \mathbf{Q} \left(\mathbf{A}_n^{(j_2)} \right)^{-1} \mathbf{X}_i^T \varepsilon_i + o_p(1) \\ &\xrightarrow{d} \mathcal{N}(\mathbf{0}, 2\sigma^2 \mathbf{Q} \mathbf{A}^{-1} \mathbf{Q}^T). \end{aligned}$$

Therefore, the second result is also proved. ■

References

- Aitkin, M. and Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society. Series B*, 47:67–75.
- Carroll, R., Ruppert, D., and Wand, M. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

- Chen, X. and Xie, M.-G. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 24:1655–1684.
- Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, 82(400):1079–1091.
- De Boor, C. (1978). *A Practical Guide to Splines*, volume 27. Springer-Verlag New York.
- Fang, Y., Lian, H., Liang, H., and Ruppert, D. (2015). Variance function additive partial linear models. *Electronic Journal of Statistics*, 9(2):2793–2827.
- Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Monographs on statistics and applied probability. Chapman and Hall, London; New York, 1st edition.
- Huang, J. (1999). Efficient estimation of the partly linear additive cox model. *The Annals of Statistics*, 27(5):1536–1563.
- Huang, J. Z. et al. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31(5):1600–1635.
- Li, R., Lin, D. K., and Li, B. (2013). Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, 29(5):399–409.
- Lin, N. and Xi, R. (2011). Aggregated estimating equation estimation. *Statistics and Its Interface*, 4(1):73–83.
- Liu, X., Wang, L., and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, 21(3):1225–1248.
- Opsomer, J. D. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics*, 8(4):715–732.
- Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, 13:689–705.

- Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, 14:590–606.
- Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.
- Wang, C., Chen, M.-H., Schifano, E., Wu, J., and Yan, J. (2015). Statistical methods and computing for big data. *arXiv preprint arXiv:1502.07989*.
- Wang, L. and Yang, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *The Annals of Statistics*, 35(6):2474–2503.
- Zhao, T., Cheng, G., and Liu, H. (2016). A partially linear framework for massive heterogeneous data. *The Annals of Statistics*, 44(4):1400–1437.